UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Identificação de perfis de consumo de bens alimentares em lares portugueses

Débora Joana Soares Alves

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:
Professora Doutora Marília Cristina de Sousa Antunes

2018

Resumo

Se antes as empresas recorriam a estratégias de marketing e vendas de massas, nas quais todos os consumidores eram tratados de igual forma, hoje mostram-se dispostas a investir muito do seu capital em estudos relacionados com o consumidor. A segmentação de mercado é uma das técnicas mais utilizadas atualmente para esse efeito. Cada vez mais, as empresas optam por esta filosofia, que divide os clientes em pequenos grupos homogéneos, *clusters*, com o objetivo de obterem vantagem relativamente à concorrência. Pela sua popularidade, acabou por torna-se na técnica *standard*, utilizada por quase todas as empresas, no mercado português de bens de grande consumo.

O estudo apresentado, neste trabalho de projeto, visa debruçar-se sobre as características sociodemográficas das famílias portuguesas e do seu cabaz de compras, por presença de produto e quantidade comprada, e analisar se estas são suficientes para segmentar e identificar perfis de consumo, no mercado de bens alimentares, através da criação de grupos de famílias, com comportamentos distintos e, consequentemente, diferentes hábitos de consumo.

Por outro lado, pretende analisar, também, associações de compra de produtos, através do estudo aos cabazes de compras de 447 famílias. Este procedimento opera numa eficaz estratégia de marketing: *cross-selling*. Esta técnica, permite encontrar padrões que ajudem à obtenção de diferentes perfis de consumo e tem como objetivo aumentar as vendas da empresa. Para isso, utiliza os resultados obtidos da análise das regras de associação, indicando os produtos comprados de forma simultânea ou produtos alternativos, quando o produto procurado não está disponível.

Os dados, para a realização desta análise, efetuada com recurso ao *software* estatístico *Rstudio*, foram fornecidos pela empresa líder de estudos de mercado, em Portugal, *The Nielsen Company*, através de uma amostra de um dos seus estudos: Painel de Lares – *CPS - Consumer Panel Services*.

Palavras-chave

Clusters, Consumo, Segmentação, Marketing.

Abstract

If before companies used marketing strategies and mass sales, in which all consumers were treated equally, today they are willing to invest much of their capital in consumer-related studies. Market segmentation is one of the most commonly used techniques for this purpose. Increasingly, companies opt for this philosophy, which divides customers into small homogeneous groups, clusters, in order to gain an advantage over competitors. Due to its popularity, it has become the standard technique, used by almost all companies, in the Portuguese market for consumer goods.

The study presented, in this project work, aims to examine the sociodemographic characteristics of Portuguese households and their shopping basket, due to the presence of product and quantity purchased, and to analyze if these are sufficient to segment and identify consumer profiles, in the food market, through the creation of groups of families, with different behaviors and, consequently, different consumption habits.

On the other hand, it also intends to analyze associations of purchase of products, through the Market Basket Analysis of 447 families. This procedure operates in an effective marketing strategy: cross-selling. This technique, widely used by manufacturers, allows to find patterns that help to obtain different consumption profiles and aims to increase sales of the company. To do this, it uses the results obtained from the analysis of association rules, indicating the products bought simultaneously or alternative products, when the product sought is not available.

The data, for this analysis, using the statistical software Rstudio, were provided by the leading market research company in Portugal, The Nielsen Company, through a sample of one of its studies: CPS - Consumer Panel Services.

Keywords

Clusters, Consumption, Segmentation, Marketing.

Índice

Resumo **iii**

Abstract **v**

Índice **vii**

Lista de Figuras **ix**

Lista de Tabelas **xi**

Lista de Gráficos **xii**

Lista de Acrónimos **xv**

Capítulo 1. Introdução **1**
1.1. Enquadramento organizacional 3
1.2. Tema de investigação 4
1.3. Objetivos do estudo 5
1.4. Relevância e justificativa 5
1.5. Motivações do estudo 6
1.6. Estrutura do documento 6

Capítulo 2. Materiais e métodos **9**
2.1. Caracterização e comportamento de compras dos lares em Portugal Continental 10
2.1.1. Evolução das tendências de consumo 10
2.1.2. O impacto das novas tecnologias no processo de decisão de compra 11
2.1.3. Atividade promocional em Portugal 12

2.2. Amostra 13
2.3. Análise de c*lusters* 14
2.3.1. Pré-processamento de dados 16
2.3.2. Medidas de proximidade 17
2.3.3. Métodos de formação de *clusters* 22
2.4. *Market Basket Analysis* 30

Capítulo 3. Análise dos dados 35
3.1. Análise exploratória 35
3.1.1. Caracterização dos lares 35
3.1.2. Caracterização das compras realizadas 39
3.2. Pré-processamento dos dados 42

Capítulo 4. Segmentação por presença de produto no cabaz de compras 47
4.1. Método de formação de *clusters* 47
4.2. Interpretação dos *clusters* 54

Capítulo 5. Segmentação por quantidade comprada no cabaz de compras 57
5.1. Método de formação de *clusters* 57
5.2. Interpretação dos *clusters* 64

Capítulo 6. Regras de associação aos cabazes de compras 69
6.1. Definição do valor dos parâmetros 69
6.2. *Itemsets* frequentes para cada nível hierárquico 71
6.3. Regras de associação para cada nível hierárquico 75
6.4. Regras de associação para *lift* inferior ou igual a 1 78

Conclusões 81

Referências Bibliográficas 85

Anexos 89

Lista de Figuras

Figura 1.1 - Diagrama com os três alvos de estudo da *Nielsen* 4
Figura 1.2 - Imagens representativas dos valores da *Nielsen* 4
Figura 2.2.1 - Identificação das áreas *Nielsen* 13
Figura 2.3.1 - Processo de criação de *clusters* 15
Figura 2.3.1.1 - Fases do pré-processamento de dados 16
Figura 2.3.2.1 - Relação entre as distâncias *Euclidiana*, *Manhattan* e *Chebyschev* 21
Figura 2.3.3.1 - Relação entre os métodos aglomerativo e divisivo 23
Figura 2.3.3.2 - Exemplo de um dendrograma 23
Figura 2.3.3.3 - Medidas de proximidade entre *clusters* 24
Figura 3.2.1 - Dados em bruto disponibilizados pela *Nielsen* 42
Figura 3.2.2 - Exemplo de classes de produtos que foram agrupadas 43
Figura 3.2.3 - Dados *raw data* em preparação 43
Figura 3.2.4 - Matriz final 43
Figura 3.2.5 - *Data frame* da matriz 44
Figura 3.2.6 - *Data frame* da matriz binária 44
Figura 4.1.1 - Dendrogramas calculados com os métodos *complete* e *single linkage* 50
Figura 4.1.2 - Índice de *Frey* - número de *clusters vs*. a função objetivo 51
Figura 4.1.3 - Índice de *McClain* - número de *clusters vs*. a função objetivo 51
Figura 4.1.4 - Índice de *C-Index* - número de *clusters vs*. a função objetivo 51
Figura 4.1.5 - Índice de *Silhouette* - número de *clusters vs*. a função objetivo 51
Figura 4.1.6 - Índice de *Dunn* - número de *clusters vs*. a função objetivo 51
Figura 4.1.7 - Índice de *PAM* - número de *clusters vs*. a função objetivo 53
Figura 4.1.8 - Dendrograma dividido em dois *clusters* 53
Figura 4.2.1 - Proporção dos *clusters*, C1 e C2, relativamente à presença de classe de produto no cabaz de compras 56
Figura 5.1 - Dendrograma utilizando a distância *Euclidiana* 60
Figura 5.2 - Dendrograma utilizando a distância *Gower* 60
Figura 5.3 - Dendrograma utilizando a distância *my.dist* 61

Figura 5.4 - Dendrograma utilizando a distância *my.dist2* 61
Figura 5.5 - Dendrograma apresentado o corte ótimo utilizando a distância *my.dist* 63
Figura 5.6 - Dendrograma apresentado o corte ótimo utilizando a distância *my.dist.2* 63
Figura 5.7 - Índice de *PAM* - número de *clusters vs*. a função objetivo aplicado à distância *my.dist* 63
Figura 5.8 - Índice de *PAM* - número de *clusters vs*. a função objetivo aplicado à distância *my.dist2* 63
Figura 6.1 - Representação de todas as CPs compradas 70
Figura 6.2 - Frequência de compra de artigos com *support* superior ou igual a 0,8 72

Lista de Tabelas

Tabela 2.3.2.1 - Tabela de contingência para variáveis binárias 19
Tabela 2.4.1 - Resumo das transações efetuadas por cinco consumidores 30
Tabela 2.4.2 - *Itemsets* frequentes com *support* mínimo de 2 31
Tabela 2.4.3 - Algumas regras de associação realizadas a partir dos *itemsets* frequentes 32
Tabela 4.1.1 - Exemplo de aplicação prática de três coeficientes 48
Tabela 4.2.1 - Peso das características sociodemográficas nos *clusters* 54
Tabela 5.1 - Resultados obtidos após aplicação do *package NbClust* 61
Tabela 5.2.1 - Resumo do número de famílias presente em cada *cluster* 64
Tabela 5.2.2 - O peso das características sociodemográficas nos *clusters* 64
Tabela 5.2.3 - Frequência média, em dias, e variedade de cabaz de compras aplicados às duas distâncias e modelos em estudo 66
Tabela 6.1 - Teste à sensibilidade do parâmetro *support* 71
Tabela 6.2 - Teste à sensibilidade do parâmetro *support* com nível de *confidence* superior a 0,75 71
Tabela 6.3 - Os 10 *itemsets* mais frequentes com nível hierárquico 1 73
Tabela 6.4 - Os 10 *itemsets* mais frequentes com nível hierárquico 2 73
Tabela 6.5 - Os 10 *itemsets* mais frequentes com nível hierárquico 3 74
Tabela 6.6 - Os 10 *itemsets* mais frequentes com nível hierárquico 4 75
Tabela 6.7 - As 10 regras de associação com nível hierárquico 2 76
Tabela 6.8 - As 10 regras de associação com nível hierárquico 3 76
Tabela 6.9 - As 10 regras de associação com nível hierárquico 4 77
Tabela 6.10 - As 10 regras de associação com valor de *lift* igual a 1 78
Tabela 6.11 - As 10 regras de associação com menor *lift* 78

Lista de Gráficos

Gráfico 3.1.1.1 - Área de residência 36
Gráfico 3.1.1.2 - Número de membros 36
Gráfico 3.1.1.3 - Tipo de família 36
Gráfico 3.1.1.4 - Classe social 36
Gráfico 3.1.1.5 - Área de residência/Número de membros 37
Gráfico 3.1.1.6 - Área de residência/Crianças 37
Gráfico 3.1.1.7 - Área de residência/Classe social 37
Gráfico 3.1.1.8 - Área de residência/Tipo de família 37
Gráfico 3.1.2.1 - Top de bebidas com maior presença nos cabazes dos lares 39
Gráfico 3.1.2.2 - Top de comidas com maior presença nos cabazes dos lares 40
Gráfico 3.1.2.3 - Média da variedade de compras realizadas 40
Gráfico 3.1.2.4 - Frequência média de compra realizada pelos lares 41
Gráfico 3.1.2.5 - Número médio de compras realizadas pelas famílias 42

Lista de Acrónimos

ROPO	Research Online, Purchase Offline
PAM	Partitioning Around Medoids
CP	Classe de Produto

Capítulo 1

Introdução

Com a forte pressão existente, pelo ganho de quota de mercado, nos bens de grande consumo, atualmente, em Portugal, pelas marcas de distribuição sobre as marcas *top of mind*, verifica-se um aumento relativamente elevado do número de estratégias aplicadas por fabricantes e retalhistas. Como forma de manterem os seus lucros e, consequentemente, o seu posicionamento no mercado, estes recorrem a técnicas de redução temporária de preço, cartões de fidelização e promoções, entre outras. Dessa forma, as marcas líderes de mercado já não possuem a mesma estabilidade de há alguns anos atrás. Vender apenas um produto de qualidade, já não é suficiente. É, sim, e cada vez mais, fundamental que as empresas tenham atenção às necessidades e comportamentos do consumidor e que, nesse sentido, as adaptem às suas ofertas.

O consumidor, agora preocupado com a relação custo/benefício dos produtos, repara e compara marcas, analisa o preço *versus* capacidade, verifica a durabilidade do produto, antecipa o consumo previsto e procura, para além de tudo isso, *feedback* de outros consumidores, para obter as informações de que necessita, na tomada de decisão. O consumidor atual é cauteloso e, contrariamente ao que sucedia, já não dá tanta relevância à publicidade presente nos meios de comunicação existentes, não se deixando, dessa forma, influenciar com tanta facilidade por anúncios, e não acredita em tudo o que vê e ouve, dando, nesse sentido, maior importância à partilha de experiências e opiniões de familiares e conhecidos.

O mercado está em constante mudança, devido a fatores económicos e financeiros, à alteração dos estilos de vida dos portugueses e à concentração de emprego nos grandes centro-urbanos, situação que acaba por criar, de forma involuntária, e cada vez mais frequente, cidades-dormitório. Desse modo, o consumidor sente a necessidade de optar por estratégias que rentabilizem, da melhor forma possível, o tempo livre de que dispõe. Assim, alguns retalhistas criaram soluções para rentabilizar o tempo dos portugueses, através, por exemplo, de mecanismos de compra online, *e-commerce*, em que a facilidade de escolha dos produtos, a encomenda e a respetiva entrega são fatores de destaque. Para além disso, dentro do supermercado, foram, também, criadas soluções eficazes, que permitem melhorar a experiência de compra do consumidor. A implementação de caixas automáticas de pagamento, que aceitam todas as modalidades de cartões e dinheiro, e a disponibilização virtual de senhas de

atendimento, fornecidas em *apps*, para os serviços de charcutaria, peixaria, entre outros, permitem ao consumidor poupar tempo durante o processo de compra.

Por outro lado, os problemas ambientais que se colocam nos dias de hoje e as abordagens regulares aos mais variados temas ligados a questões de saúde das populações são uma constante na vida dos portugueses. Por isso, a preocupação com o tipo de ingredientes que compõem os produtos aumenta, alterando, assim, o conceito de produto *premium*, que é agora considerado um produto fabricado com ingredientes 100% naturais. Os produtos sem lactose ou glúten, por exemplo, integram também esta gama, tendo aumentado a sua quantidade e diversidade no mercado de bens de grande consumo.

Torna-se, assim, pertinente estudar o comportamento do consumidor português, das suas necessidades e expectativas, através da análise da frequência e do local onde faz as suas compras. Esta investigação permite fornecer não só uma análise detalhada daquilo que é o mercado de bens de consumo, tendo em conta a amostra utilizada, mas também demonstrar as diferentes oportunidades de mercado que existem, que ainda não estão totalmente desenvolvidas e que dependem da atenção que marcas e retalhistas dão ao consumidor, que é, hoje, mais exigente nas suas escolhas. Para isso, são necessárias estratégias e técnicas cada vez mais sofisticadas, que permitam entender, com rigor, o comportamento do consumidor, de forma a minimizar os prejuízos provocados por crises emergentes.

Mais do que nunca, as empresas mostram-se dispostas a investir muito do seu capital, em estudos relacionados com o consumidor, com o objetivo de obterem vantagem relativamente à concorrência. Uma das técnicas atuais mais utilizadas é a segmentação de mercado, que, pela sua popularidade, acabou por tornar-se na técnica *standard*, utilizada por quase todas as empresas, no mercado de bens de grande consumo. Porém, ao estudar o consumidor na ótica da segmentação, torna-se necessário estabelecer, *a priori*, o tipo de segmentação pretendido e definir qual o seu objetivo. Para Kotler (2000), por forma a dividir o consumidor em grupos homogéneos, é necessário estabelecer o tipo de segmentação adequada. Segundo o autor, existem quatro tipos de segmentação de mercado: geográfica, demográfica, comportamental e psicográfica/psicológica. Uma segmentação de mercado visa dividir o mercado em grupos homogéneos, em função das suas necessidades e expectativas. A cada segmento associam-se diferentes hábitos e produtos comprados, e, possivelmente, cada grupo requer necessidades distintas, sendo que as marcas têm de conseguir dar resposta a toda esta diversidade ou, dependendo da dimensão do grupo em questão, optar por não abranger todo o mercado – nichos de mercado.

O estudo apresentado, neste trabalho de projeto, procura abordar duas análises. A primeira é referente às características sociodemográficas das famílias portuguesas e do seu cabaz de compras, por presença de produto e quantidade comprada, e permite analisar se estas características são suficientes para segmentar e identificar perfis de consumo, no mercado de bens de grande consumo, em Portugal, através da criação de grupos de famílias, *clusters*, com comportamentos distintos e, consequentemente, diferentes hábitos de consumo.

Por norma, são utilizadas técnicas de análise multivariada para agrupar, neste caso, famílias, em segmentos de mercado. Apesar da diversidade de técnicas existentes – análise de classes latentes, redes neuronais ou árvores de decisão, entre outras –, a análise de *clusters* continua a ser o procedimento mais comum e largamente aplicado, quando o objetivo é a segmentação de mercado.

As técnicas de extração de conhecimento de dados têm sido utilizadas com sucesso num grande número de problemas reais (Gama, Carvalho, Faceli, Lorena & Oliveira, 2012). Dessa forma, a segunda análise incide na técnica *market basket analysis*, com foco em regras de associação, que opera de forma bastante

eficaz numa estratégia de marketing conhecida e muito utilizada: *cross-selling*. Considera-se que a utilização desta técnica se traduz na criação de *insights* importantes, para a tomada de decisão das empresas.

Sendo considerada uma das áreas mais antigas de *data mining*, o *market basket analysis* procura aplicar as relações relevantes encontradas em bases de dados de transações de retalho, com o objetivo de potenciar vendas e atrair clientes (Raeder & Chawla, 2011). Dessa forma, são estudadas as associações aos produtos presentes nos cabazes de compras das 447 famílias analisadas. Esta técnica utiliza algoritmos de regras de associação, que procuram investigar a correlação ou associação entre conjuntos de artigos comprados com maior frequência, indicando os produtos comprados de forma simultânea ou produtos alternativos, quando o produto procurado não está disponível.

Assim, a *market basket analysis* constitui uma ferramenta extremamente importante no sistema de retalho organizacional, focando-se nos cabazes de consumo dos clientes para monitorizar padrões de compra e potenciar a satisfação do cliente (Microstrategy, 2003).

Os dados, para a realização desta análise, efetuada com recurso ao *software* estatístico *Rstudio*, foram fornecidos pela empresa líder de estudos de mercado em Portugal, *The Nielsen Company*, através de uma amostra de um dos seus estudos: Painel de Lares – *CPS - Consumer Panel Services*.

1.1. Enquadramento organizacional

The Nielsen Company é uma empresa que estuda consumidores, em 47 mercados europeus e em mais de 100 países em todo o mundo, com o objetivo de proporcionar uma visão completa das tendências de consumo. Fundada em 1923, nos Estados Unidos da América, e presente em Portugal desde 1967, a *Nielsen* já atingiu a sua fase de maturidade, sendo atualmente líder nacional e mundial em estudos de mercado.

No caso português, a empresa apresenta três alvos de estudo, ilustrados na Figura 1.1: Estudo ao Consumidor (*CI - Consumer Insights*), Painel de Lares (*CPS - Consumer Panel Services*) e Painel de Retalho (*RMS - Retail Measurement Services*). É, neste último caso, a única a trabalhar, a nível de painéis de retalho de bens de grande consumo, em todo o país. Mantém, também, 12 estudos de mercado em continuidade, cada um com um objetivo, adaptando-se aos diferentes tipos de cliente e procurando ir ao encontro das suas necessidades.

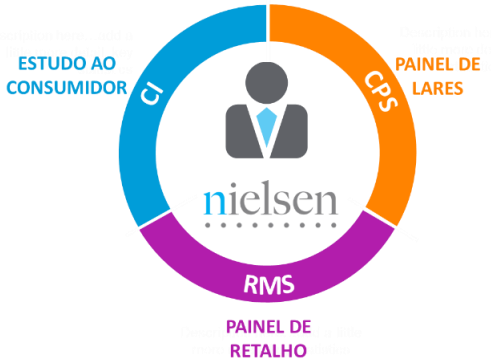

Figura 1.1 - Diagrama com os três alvos de estudo da *Nielsen*

The Nielsen Company tem como missão o contributo para o êxito dos seus clientes, proporcionando-lhes um melhor conhecimento do mercado, ao mesmo tempo que ambiciona ser reconhecida mundialmente como a melhor empresa de estudos de mercado. Recentemente, renovou os valores que procura transmitir tanto interna, como externamente.

Figura 1.2 - Imagens representativas dos valores da *Nielsen*

Nesse sentido, e como demonstra a Figura 1.2., a *Nielsen* pretende ser uma empresa *Open*, ou seja, aberta a novas ideias e a uma perspetiva *think outside the box*; *Connected*, do ponto de vista da eficiência e eficácia, a empresa estará sempre conectada e alinhada tanto internamente, como com as prioridades do cliente; *Useful*, tornando-se ainda mais útil, de modo a conseguir ir ao encontro das necessidades do cliente – *the more useful, the more valuable*; e, por fim, *Personal*, procurando desenvolver o lado mais pessoal do seus colaboradores e das relações com o cliente.

1.2. Tema de investigação

Este trabalho insere-se no tema *Data Mining* e Identificação de Padrões, através, nomeadamente, da identificação de perfis de consumo de bens alimentares, em lares portugueses, em Portugal Continental.

Os dados recebidos foram objeto de dois estudos distintos: análise de *clusters* e *market basket analysis*.

Na primeira análise, foram agrupados agregados familiares, de acordo com os seus hábitos de compra, segmentando-os por presença e variedade de CP. Para isso, recorreu-se a indicadores socioeconómicos, área geográfica, número de membros dos agregados analisados, entre outros.

A segunda análise, *market basket analysis*, com foco em regras de associação, procura investigar a correlação ou associação entre conjuntos de artigos comprados com maior frequência, de modo identificar padrões de consumo.

1.3. Objetivos de estudo

Este estudo tem como objetivos gerais:
(i) Segmentação de famílias portuguesas, de acordo com as compras efetuadas por cada um dos lares em análise, durante um período de três meses e meio, através de metodologias de *clustering*;
(ii) Utilização de regras de associação, para identificação de grupos de produtos potencialmente comprados em conjunto, nos diferentes perfis de consumo.

1.4. Relevância e justificativa

A identificação de perfis de consumo de bens alimentares em lares constitui uma mais-valia para o conhecimento do comportamento alimentar das famílias portuguesas. Este conhecimento é importante e interessante tanto para as áreas de Marketing, que podem assim direcionar de forma mais eficiente as campanhas publicitárias, como para a área da Saúde, como suporte a medidas de saúde pública, uma vez que o consumo alimentar está intimamente relacionado com a saúde das populações.

Com as oscilações presentes atualmente no mercado, é cada vez mais importante estudar o comportamento de cada consumidor. Desse modo, torna-se necessário compreender o que leva os consumidores a terem determinadas escolhas e, com isso, identificar de que forma os fatores financeiros afetam o cabaz de compras de cada lar.

A concorrência existente entre as empresas é cada vez maior, da mesma forma que os métodos utilizados para o incentivo à compra são cada vez mais vastos – desde promoções, descontos em talão e cartão, compras leve 2 pague 1, entre outras –, tendo estes cada vez menos impacto no consumidor. Mais do que nunca, as empresas estão dispostas a investir muito do seu capital neste tipo de estudos, por forma a ganharem posição relativamente à concorrência (Armstrong & Kotler, 2011).

Uma forma de minimizar estes obstáculos, passa pela utilização de técnicas que têm vindo a ganhar cada vez mais importância no setor, com destaque para a compra de estudos a painel de lares (CPS) ou mesmo estudos ao consumidor (CI). Outra das soluções utilizadas pelas empresas, de forma a direcionarem as suas campanhas publicitárias ou a divulgação de um novo produto, por exemplo, é a segmentação de mercado. Mas não só. De acordo com Salvador e Campomar (2014), toda a estratégia de marketing é baseada em segmentação, seleção de mercado-alvo e posicionamento.

1.5. Motivações do estudo

Ao longo dos anos, o consumidor português tem vindo a alterar os seus comportamentos, situação que modificou as tendências de consumo, a todos os níveis, e criou implicações nas empresas, como referido anteriormente, obrigando-as a alterar a sua forma de comunicar com os consumidores. Tornou-se, desse modo, essencial analisar, em detalhe, os hábitos de consumo dos portugueses, para apurar o que gerou estas alterações, mas também o que poderá ser feito, no futuro, para acompanhar esta evolução. Desse modo, procurou dar-se resposta, através do estudo ao painel de lares, a várias questões sobre o consumidor de hoje em dia e os diferentes hábitos alimentares existentes, por forma a identificar perfis de consumo, em lares portugueses.

Considerou-se, por isso, uma vantagem, em relação a outros estudos já existentes, o facto de a análise ser realizada a partir de dados fidedignos e reais da empresa líder de mercado, em Portugal. Os dados foram recebidos em bruto, sem serem trabalhados, não condicionado, em nada, as conclusões que daí se podiam extrair, ao contrário de outros trabalhos, desenvolvidos por empresas ou investigadores, a partir de questionários, que, de certa forma, não conseguem controlar os métodos e técnicas de pesquisa utilizados. Por outro lado, os questionários são realizados com base na memória do consumidor, enquanto os dados utilizados neste trabalho são recolhidos a partir de ações reais – declarado *versus* real.

No que diz respeito à amostra, esta foi selecionada aleatoriamente pela *Nielsen*, sendo constituída por 447 lares. A dimensão amostral, assim como os critérios de seleção da mesma, foram definidos pela própria empresa, com o intuito de salvaguardar o anonimato dos dados. No entanto, considera-se que a dimensão amostral é robusta dado que, para uma amostra de 447 lares, o erro máximo associado, para um nível de confiança de 95%, é de aproximadamente 5%.

O painel é por si próprio representativo da população portuguesa, uma vez que abrange todos os tipos de famílias, de todas as faixas etárias.

Em suma, este trabalho, visa, para além dos objetivos principais já indicados, preencher algumas lacunas de outras investigações da mesma área.

1.6. Estrutura do documento

A presente investigação divide-se em duas partes: dois capítulos de revisão teórica, para aprofundar o tema em análise, e outros três de estudo empírico, incluindo a metodologia utilizada.

No atual capítulo, são definidos os objetivos da investigação, é feito o enquadramento organizacional sobre a *Nielsen*, empresa na qual o trabalho de projeto foi realizado, e, igualmente, sobre o tema de investigação, abordando a sua relevância e justificativa e as motivações técnicas do trabalho.

No segundo capítulo, Materiais e Métodos, é apresentada uma breve caracterização do comportamento dos lares, em Portugal Continental, através de três temas importantes: evolução das tendências de

consumo, impacto das novas tecnologias no processo de decisão de compra e atividade promocional. É, também, descrita a amostra e exposto o processo de recolha de informação. Por fim, são ainda apresentados outros dois subcapítulos. Um sobre a análise de *clusters*, onde são definidas as medidas de proximidade e os métodos hierárquicos e não hierárquicos, e outro sobre *market basket analysis*, com especial foco nas regras de associação.

O terceiro capítulo introduz a amostra, através da análise exploratória, com a caracterização dos consumidores e das compras realizadas, e apresenta o pré-processamento dos dados.

Nos capítulos 4 e 5, procedeu-se à análise concreta dos dados, recorrendo à segmentação por presença de produto e quantidade comprada, no cabaz de compras de cada lar, através do processo de formação de *clusters*. Para isso, foi analisada a viabilidade dos métodos descritos na revisão teórica, com recurso a *software* estatístico apropriado – *Rstudio*. Por fim, é realizada a interpretação dos resultados obtidos, através da análise de semelhança entre os *clusters*.

No capítulo 6, foi realizada a análise *market basket analysis*. Desse modo, recorreu-se aos cabazes de compras das famílias portuguesas, com o objetivo de encontrar padrões de consumo, ou seja, correlações ou associações de produtos que são comprados em simultâneo, produtos complementares ou produtos substitutos[1].

Por último, são apresentadas as conclusões do trabalho, as referências bibliográficas e os anexos, com informação complementar sobre as famílias, nomeadamente as definições dos tipos de famílias existentes, e as áreas geográficas *Nielsen*.

[1] Produtos alternativos quando o produto procurado inicialmente não está disponível.

Capítulo 2

Materiais e Métodos

Em 2004, o estudo de painel de lares, em Portugal Continental, passou a ser produzido pela *Nielsen*. Atualmente, a empresa detém uma amostra de 3 mil lares, representativa da população portuguesa. Para isso, utiliza tecnologia *scanning*[2], que permite recolher continuamente o histórico de compras diárias de bens de grande consumo. Dessa forma, a informação reunida permite compreender o comportamento dos consumidores, de modo a garantir o sucesso de vendas a longo prazo.

Para a realização do presente estudo, foi utilizada uma amostra dos 3 mil lares. Os dados foram agregados e anonimizados, não existindo, por isso, forma de os relacionar com o lar a que pertencem.

Nesse sentido, é apresentado um subcapítulo teórico, onde são definidas as medidas de proximidade e os métodos hierárquicos e não hierárquicos, que serão utilizados na componente empírica do trabalho, para a segmentação da amostra recebida, através de uma análise de *clusters*. Dessa forma, explica-se em detalhe todo o processo de formação de *clusters*, com recurso ao *software* estatístico *RStudio*, versão 1.0.153, através da aplicação de *packages* específicos para o efeito.

[2] Tecnologia que reúne e compila as informações sobre o cabaz de compras de cada lar, a partir do processo de leitura dos códigos de barras dos produtos comprados, através de um dispositivo próprio utilizado para o efeito.

2.1. Caracterização e comportamento de compras dos lares em Portugal Continental

> "Há um ano atrás, a situação política nacional era mais instável e a incerteza tomava conta das principais preocupações do consumidor português. Agora mais confiantes, os portugueses passam a preocupar-se especialmente com questões mais pessoais, nomeadamente o equilíbrio entre a vida pessoal e profissional, a saúde e a família" (Barbosa, 2017 citado por Nielsen Portugal, 2017a).

2.1.1. Evolução das tendências de consumo

A tendência de consumo tem sido um tema em destaque nos últimos anos, devido à divergência de definições relativamente ao tipo de consumidor existente, atualmente, em Portugal. Se muitos consideram que o consumidor continua a estar dividido em grupos homogéneos, outros defendem que os hábitos se alteraram, tornando mais difícil a criação de *clusters* populacionais, como forma de responder às necessidades do grupo.

Recentemente, tem-se verificado uma melhoria no consumo, em Portugal, resultado da diminuição da taxa de desemprego e do consequente aumento do poder de compra. Ainda existe, contudo, 18% da população portuguesa que afirma não lhe sobrar dinheiro depois de pagar as despesas mensais. Apesar disso, um terço admite que a sua situação financeira melhorou nos últimos cinco anos (Barbosa, 2017).

Hoje, um maior acesso à informação, por parte dos consumidores, é um dos principais fatores para a criação de grupos heterogéneos. Porém, quando se aborda a categoria de bens tecnológicos, talvez essa regra não se aplique, devido, entre outros fatores, à grande notoriedade de algumas marcas no mercado. Assim, ao ter valor percecionado de uma marca, a associação a um determinado estatuto social desejado, talvez seja um fator que está presente aquando da decisão de compra dos consumidores.

O acesso a uma maior informação por parte do consumidor alterou também o tipo de produtos comprados. Este mostra-se, agora, mais consciente e preocupado com a saúde, ao dar mais atenção aos produtos alimentares adquiridos. Esse é, aliás, um indicador com valores superiores à média europeia, com 66% dos consumidores portugueses a afirmarem estar dispostos a pagar mais por produtos alimentares e bebidas que não contenham substâncias artificiais, acreditando que o corte nestes alimentos será benéfico para a sua saúde e a daqueles que os rodeiam (Barbosa, 2017).
Deste modo, a definição de produto *premium*[3] também se alterou, considerando-se agora produtos com qualidade superior em matérias-primas ou ingredientes 100% naturais. Com isso, o valor dos bens acaba por ser inflacionado, refletindo-se, mais tarde, no custo final para os compradores. Tendo em conta a média europeia, 38% dos consumidores consideram que os preços altos tornam o produto mais *premium*, fator que apenas é valorizado por 21% dos portugueses (Nielsen Portugal, 2016a).

Categorias como Carne e Peixe, 43%, estão no topo das preferências *premium*, seguindo-se Vestuário e Calçado, com 37%. Por outro lado, Produtos de Papel, 4%, e Snacks Salgados, 4%, são as categorias

[3] Considera-se *premium* um produto composto por ingredientes únicos e de alta qualidade e com preço bastante elevado.

para as quais os consumidores portugueses estão menos dispostos a pagar preços mais elevados (Nielsen Portugal, 2016a).

O ritmo de vida acelerado dos consumidores tem, igualmente, impulsionado a procura por produtos e serviços mais práticos e flexíveis. Nesse sentido, tudo o que possa simplificar a vida do consumidor terá potencial para crescer: refeições prontas a comer, *take away* e produtos de IV gama[4] (Barbosa, 2017).

2.1.2. O impacto das novas tecnologias no processo de decisão de compra

Com o desenvolvimento da tecnologia, foram várias as alterações nos hábitos dos consumidores portugueses. Atualmente, o consumidor consegue obter informações sobre um determinado produto sem sair de casa – situação impensável há uma década atrás. A nova tendência de decisão de compra passa pela *ROPO*[5], onde o consumidor pesquisa *online* informações relevantes sobre o produto que pretende comprar, deslocando-se, depois, a uma loja física para o fazer. Nos setores Vestuário e Eletrónica, os consumidores apoiam-se na pesquisa em lojas físicas, lojas *online* e sites das lojas. No entanto, na categoria Beleza e Cuidado Pessoal, os portugueses procuram informação em lojas físicas, em *websites* com cupões e descontos e, ainda, a partir das recomendações dos seus conhecidos (Nielsen Portugal, 2017b).

Com o crescimento dos blogues e das redes sociais, fruto do desenvolvimento da web 2.0, o consumidor passou a ter cada vez mais informação disponível sobre os produtos e deixou de ser passivo no processo de compra. A partilha de experiências e opiniões influencia, igualmente, e cada vez mais, as decisões de compra de outros consumidores.

Atualmente, e apesar de o *e-commerce* ainda estar pouco desenvolvido em Portugal, uma vez que os mecanismos de compra *online* ainda não estão bem definidos, o consumidor português mostra-se, comparativamente aos europeus, mais disponível para a realização de compra no ciberespaço. A facilidade de pagamento e a entrega do produto são alguns dos fatores que o atraem. Assim, 61% dos portugueses admite sentir-se seguro na disponibilização de informações pessoais nestas plataformas. Existem, porém, algumas categorias – como Frescos ou Mercearia – pelas quais 59% dos consumidores se revelam menos disponíveis para comprar *online*, preferindo fazê-lo em lojas físicas, de forma a serem eles próprios a escolher os produtos. Para além disso, os custos adicionais de entrega são apontados como sendo, igualmente, um fator de destaque para que a maioria dos portugueses não considere conveniente esta modalidade de compra (Nielsen Portugal, 2017b).

De facto, o desenvolvimento das novas tecnologias veio facilitar bastante a vida dos consumidores, dando-lhes agora mais tempo para realizarem outras tarefas. Essa é uma situação apenas possível graças à vantagem de se poder comprar o que quiser, quando quiser e onde quiser, não existindo a necessidade de deslocação física aos estabelecimentos comerciais. Desse modo, o processo de compra de qualquer produto tornou-se fácil, sem demoras ou filas de espera, podendo ser efetuado através de computador,

[4] Definem-se como produtos lavados, embalados e prontos a consumir ou confecionar.
[5] Traduz-se pela pesquisa *online* antes da compra no local.

tablet ou *smartphone*[6], bastando para isso acesso à Internet. Assim, tudo o que rentabilize o tempo do consumidor e lhe traga conveniência no processo de compra, terá tendência para crescer.

No que toca às compras em loja, as alternativas, hoje disponíveis, também se alteraram. Para além de ser possível a compra *online* e o respetivo levantamento em loja, a introdução de caixas *self-service* e de *scanners* tem tornado a experiência de compra em loja mais atrativa. Em Portugal, 41% dos consumidores admite utilizar as caixas *self-service*, disponíveis nos vários estabelecimentos comerciais, como forma de poupar tempo. Além disso, 70% mostram-se disponíveis para utilizar *scanners* manuais, que permitem uma compra autónoma, de forma a evitar filas nas caixas de saída (Barbosa, 2017).

Para além disso, a implementação de *apps* veio permitir ao cliente, entre outras coisas, tirar senha para os serviços de charcutaria, peixaria ou padaria. Estas soluções tornam possível o acompanhamento, através de *tablet* ou *smartphone*, do número que está a ser atendido, permitindo ao cliente rentabilizar o seu tempo nas restantes compras, e ajudam a evitar aglomerados de consumidores junto desses serviços.

2.1.3. Atividade promocional em Portugal

Desde 2011, a atividade promocional duplicou em Portugal, levando a uma valorização acima da média das promoções e dos preços baixos (Nielsen Portugal, 2016b). Os descontos são cada vez maiores, mas a eficiência gerada pelas promoções de 50%, por exemplo, é cada vez menor. Isto significa que, apesar de as promoções serem importantes, começa a notar-se saturação em algumas categorias.

Atualmente, existem vários fabricantes a vender os seus produtos abaixo do preço de custo (Silva, 2016), de forma a conseguirem manter-se competitivos face às marcas de retalhista, preferindo, como consequência dessa política, o pagamento de coimas à perda de quota de mercado. Todavia, a redução temporária de preço é talvez o método mais utilizado para estimular a competitividade neste mercado, por conseguir resultados a curto prazo.

Existem inúmeras técnicas que podem ser utilizadas para promover o aumento de vendas. As mais comuns são: bónus *pack*, em casos onde exista uma quantidade grátis do mesmo produto; multi-compra, uma promoção imediata obtida pela compra de mais do que uma unidade de produto; o clássico leve 3 pague 2; e cartões de lealdade/fidelização.

Nos anos 90, como forma dos fabricantes aumentarem as suas vendas, alguns produtos, como bolos ou batatas fritas, continham brindes nas embalagens, para atrair um público mais jovem. A maioria dessas ofertas contemplava uma coleção de artigos, o que obrigava a uma fidelização e a um consequente período de consumo continuado por parte dos clientes. No entanto, as estratégias das marcas têm evoluído bastante no tipo de técnicas utilizadas e a sua criatividade não tem limites, embora atualmente estas ofertas de brindes não tenham grande impacto, uma vez que o consumidor é mais consciente e entende que a aquisição do produto, tendo por base a oferta, não lhe acrescenta qualquer benefício.

[6] Para além do processo de compra convencional *online*, existem ainda *apps*, disponíveis para as várias plataformas móveis, para utilização personalizada em *tablet* e *smartphone*.

Recentemente, a ação *Gang dos Frescos*[7], promovida pelo retalhista *LIDL*, associado à sensibilização para a adoção de hábitos alimentares saudáveis, através do consumo de frutas e legumes, foi um caso de sucesso (Activa, 2015). Nesta situação, existia um propósito social, que pode ter sido determinante para as proporções que a campanha atingiu junto dos consumidores. Para confirmar esse sucesso, basta verificar que o método utilizado nessa ação foi seguido por outros retalhistas nacionais.

As marcas estão demasiado competitivas entre si, não dando atenção ao mais importante: satisfazer as necessidades do cliente – cerca de 81% dos portugueses gostariam de ter ofertas promocionais personalizadas (Barbosa, 2017). Atualmente, já é possível, através do histórico de compras presente em cada cartão de lealdade/fidelização, tentar satisfazer as expectativas do consumidor. Para isso, "é essencial que as marcas e insígnias conheçam muito bem os seus consumidores: perfil, comportamento de compra e expectativas" (Barbosa, 2017, p. 37). A grande tendência passa, assim, por adequar as ofertas promocionais personalizadas, como forma de atrair o cliente.

2.2. Amostra

Para este estudo, foram utilizados dados relativos a três meses do *raw data*[8] de compras, correspondentes ao período compreendido entre 4 de janeiro e 17 de abril de 2016, com a identificação da hora. A amostra corresponde a 447 lares – cerca de 15% dos 3 mil que cooperam com a Nielsen –, com o corresponde *ID*[9] de cada um deles, e está repartida pelas diferentes áreas *Nielsen*[10], ilustradas na Figura 2.2.1.

Figura 2.2.1 - Identificação das áreas *Nielsen*

[7] A primeira campanha foi realizada em 2014. Esta ação visou promover o aumento da frequência de ida ao supermercado, mas também o aumento dos cabazes de compras, uma vez que a partir de compras iguais ou superiores a 10 euros, o cliente recebia uma saqueta com 4 cartas e um selo para colecionar. A partir do momento em que reunia 6 ou 12 selos, era possível adquirir um dos peluches disponíveis – cogumelo, figo, maçã, ervilha, tomate, pimento amarelo, laranja e mirtilo – por 9,99 ou 2,99 euros, respetivamente.

[8] *Raw data* significa dados primários. Representa um conjunto de dados recolhidos em bruto.

[9] *ID* significa *Identity*. Traduz-se por um valor único atribuído a um objeto, de modo a identificá-lo.

[10] A empresa divide Portugal Continental em 6 áreas: I (Grande Lisboa), II (Grande Porto), III Norte (Litoral Norte), III Sul (Litoral Sul), IV (Interior Norte) e V (Sul). As regiões autónomas dos Açores e da Madeira estão fora da amostra (ver Anexo A).

O estudo de lares, realizado pela multinacional, inclui três grandes grupos: Comida, Bebida e Drogaria. De forma a simplificar a análise, apenas foram utilizadas as indústrias *Food and Beverages*, divididas por 181 classes de produto[11].

No documento *raw data* disponibilizado, foram incluídas variáveis quantitativas – número de membros do agregado familiar – e qualitativas – classe social, tipo de família e presença de crianças no lar (ver Anexo B).

Processo de recolha de informação

Os dados dos compradores são recolhidos por *scanners* manuais e transmitidos para a *Nielsen*, uma vez por semana, de forma automática, não tendo esta transmissão qualquer custo associado para o lar. Esta tecnologia de ponto de venda, para serviços de medição de venda a retalho, capta os dados de vendas e de preços de praticamente todas as grandes cadeias de retalho. Assim, o lar, ao tornar-se membro do painel, terá obrigatoriamente de registar todas as compras que efetuar, com o auxílio do leitor de códigos de barras fornecido pela *Nielsen*. Para isso, deverá fazer a respetiva leitura do código de barras de cada produto que comprou.

Pelo envio da informação são recebidos pontos de recompensa, que poderão ser trocados por uma variedade de prémios, que constam num catálogo de ofertas exclusivas. A informação recebida de cada membro é compilada, por aglomerado, de forma anónima.

Posteriormente, estes dados são utilizados em vários relatórios e serviços da *Nielsen*, vendidos a fabricantes e retalhistas, de modo a melhorar os seus produtos e serviços.

2.3. Análise de *Clusters*

Nos últimos anos, as técnicas de análise de *clusters* têm vindo a desenvolver-se com o intuito de possibilitar investigações. Este tipo de estudos é bastante útil para efeitos de marketing, na medida em que permite identificar o perfil de cada consumidor e segmentá-lo, por forma a oferecer um tratamento mais personalizado e de acordo com os seus interesses. A segmentação de mercado, no geral, é, de facto, uma área que recorre frequentemente à análise de *clusters* (Hand, Mannila & Smyth, 2001).

A análise de *clusters* é uma técnica exploratória da análise multivariada, que tem como principal objetivo a construção de grupos ou *clusters* homogéneos. Este agrupamento é efetuado para que elementos pertencentes ao mesmo grupo tenham características semelhantes e elementos de diferentes grupos tenham características dissemelhantes. De um modo geral, pretende dividir-se o conjunto de indivíduos em grupos, onde os membros de um mesmo grupo são mais próximos entre si do que dos membros de outros grupos. Para a obtenção destes grupos, é necessário utilizar uma medida de proximidade entre indivíduos, mas também entre grupos de indivíduos. Estes indivíduos são caracterizados por variáveis

[11] Agrupamento de produtos semelhantes, dentro da mesma categoria.

de natureza qualitativa e/ou quantitativa, que têm de ser tidas em conta para o cálculo destas medidas de proximidade.

Para a realização da análise de *clusters*, é necessário tomar decisões e ter em conta aspetos que dependem de cada caso em particular. Contudo, existe um conjunto de passos *standard*, para todos os casos. Genericamente, segundo Reis (2001), a análise de *clusters* compreende cinco etapas:
1. Seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. Definição de um conjunto de variáveis, a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. Definição de uma medida de semelhança ou distância entre cada dois indivíduos;
4. Escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição de um algoritmo de partição/classificação;
5. Validação dos resultados obtidos.

Também Branco (2004) apresenta um processo de criação de *clusters*, recorrendo, contrariamente a Reis (2001), a uma etapa facultativa de transformação de variáveis, como demonstra a Figura 2.3.1.

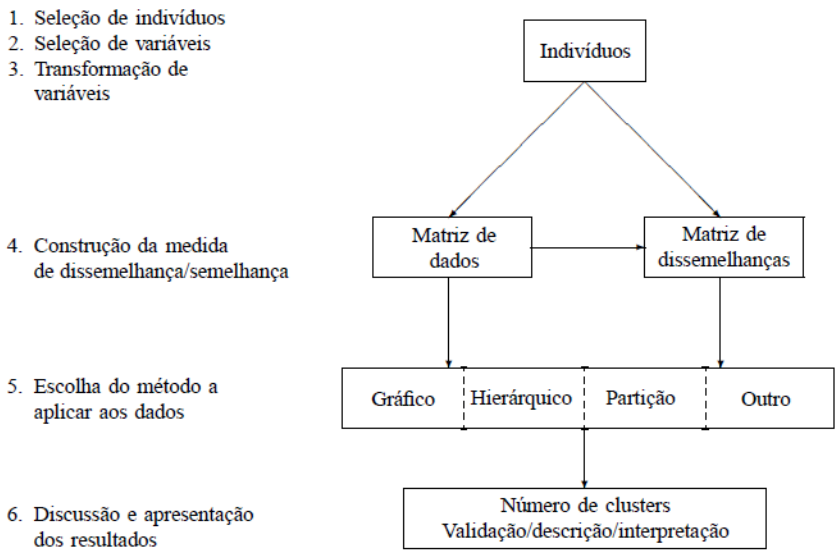

Figura 2.3.1 - Processo de criação de *clusters* (Branco, 2004)

Estas etapas, segundo Branco (2004), têm como objetivo responder a um conjunto de questões que se colocam no decorrer da análise, das quais se destacam:
1. Que indivíduos se pretende agrupar?
2. Que variáveis se devem considerar para caracterizar esses indivíduos?
3. Existem valores omissos ou valores errados que podem ser corrigidos? Como integrar informação sobre os dados recolhidos em fontes distintas? Todas as variáveis são relevantes para a análise?
4. Devem as variáveis ser transformadas de alguma maneira? Em resumo, que tipo de pré-processamento deve ser aplicado aos dados antes de os agrupar?
5. Qual a medida de proximidade que deve ser utilizada entre indivíduos?
6. Dos métodos de formação de *clusters*, por qual deve optar-se tendo em vista o problema em particular?
7. O que distingue um *cluster* de outro? Qual a maneira mais clara e sucinta de sumariar os resultados e como validá-los?

2.3.1. Pré-processamento de dados

De modo a poder iniciar-se o processo de análise de *clusters*, é necessário trabalhar os dados, com o objetivo de os organizar o melhor possível e de forma mais relevante, para a análise. A esse conjunto de procedimentos dá-se o nome de pré-processamento de dados.

Como em todos os processos, também aqui existe um conjunto de etapas a seguir. O pré-processamento de dados em análise, apresentado por Han, Kamber e Pei (2011), é efetuado em quatro fases, representadas na Figura 2.3.1.1.

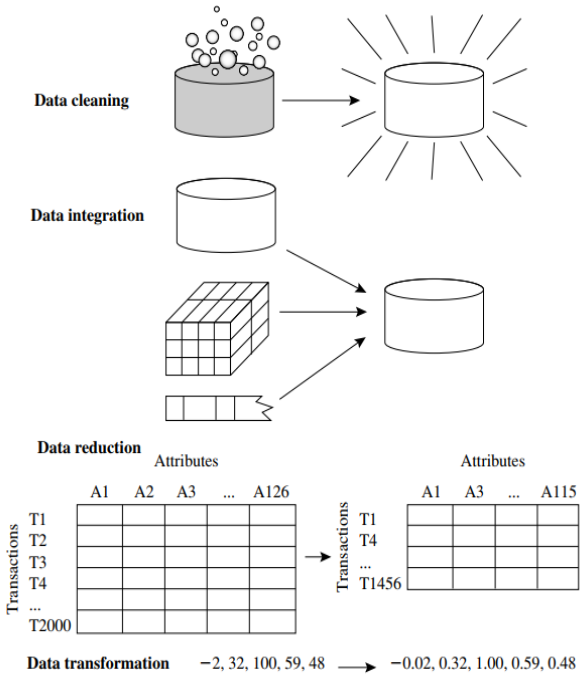

Figura 2.3.1.1 - Fases de pré-processamento de dados (Han *et al*., 2011)

Dados reais tendem a ser incompletos, inconsistentes e ruidosos[12]. Desta forma, a primeira fase inicia-se pelo *data cleaning* – limpeza dos dados através da remoção do ruído ou da correção de alguns valores omissos ou incongruentes com os restantes dados.

De seguida, procede-se à *data integration*. Como o próprio nome indica, esta etapa trata da integração dos dados, através da fusão de várias fontes. Uma integração cuidadosa pode ajudar a reduzir e a evitar redundâncias e inconsistências no conjunto de dados, o que pode ajudar, nalguns casos, a melhorar a precisão e a velocidade do processo de *data mining*.

Com análises complexas, de grande dimensão, que podem levar bastante tempo a ser processadas, tornando a análise impraticável ou inviável, é necessário ter em atenção o passo três – *data reduction*. Este passa pela transformação das variáveis, de modo a reduzir a dimensão do conjunto de dados, mantendo, porém, a sua integridade.

[12] O ruído nos dados é um erro ou variância aleatória numa variável.

Por último, *data transformation* e *data discretization* tratam da transformação e da consolidação das variáveis, para que possam melhorar o desempenho dos algoritmos envolvidos, tornando o processo de *data mining* mais eficiente.

É certo que o analista terá de verificar todo o processo de pré-processamento, para identificar quais os passos mais relevantes que devem ser considerados, de acordo com os dados que tem disponíveis e do próprio objetivo da análise, de forma a aplicar os procedimentos mais adequados.

2.3.2. Medidas de proximidade

A escolha das medidas de proximidade é de extrema importância na análise de *clusters*, pois dela depende a identificação dos indivíduos que estão mais próximos ou mais distantes um do outro, dependendo das características suscetíveis de análise.

Considera-se, desse modo, que os dados a analisar são constituídos por n objetos, podendo estes representar indivíduos, produtos, lojas, entre outros. Os algoritmos da análise de *clusters* podem ser tipicamente aplicados a duas estruturas de dados, de acordo com Han *et al*. (2011).

1. Matriz de dados ou estrutura indivíduo-por-variável: Usualmente representada por X, a matriz de dados tem, genericamente, dimensão $n \times p$, em que cada linha contém a informação referente a um indivíduo, medido em p variáveis, $X_1, X_2, \cdots, X_p$, que os caracterizam.
Considera-se um conjunto de n indivíduos e p variáveis, dispostos na seguinte matriz de ordem $n \times p$,

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

O elemento x_{ij} representa o valor da variável X_j no indivíduo i.

2. Matriz de dissemelhanças ou estrutura indivíduo-por-indivíduo: É uma matriz de dimensão $n \times n$, cujas entradas correspondem às dissemelhanças entre cada par dos n indivíduos. Entende-se por dissemelhança entre dois indivíduos uma medida que reflete a diferença entre esses indivíduos.

Seja D uma matriz de ordem n, dada por $D = \begin{bmatrix} 0 & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & 0 \end{bmatrix}$, em que d_{kl} representa a distância entre o indivíduo k e l, com $k, l = 1, \ldots, n$. Esta matriz designa-se por matriz de proximidade.

Para a sua construção, é necessário selecionar a medida de proximidade a utilizar, podendo esta ser uma medida de semelhança ou dissemelhança/distância. É, por isso, necessário escolher a medida de proximidade que melhor se adequa ao tipo de dados da amostra.

É possível caracterizar a medida de dissemelhança, entre indivíduos i e j, δ_{ij}, através das seguintes propriedades:

1) $\delta_{ij} \geq 0$: A medida é não negativa.
2) $\delta_{ii} = 0$: A medida de um elemento a si próprio é zero.
3) $\delta_{ij} = \delta_{ji}$: A medida entre dois elementos é simétrica.
4) $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$: Conhecida por desigualdade triangular, especifica que a menor distância entre dois pontos é a direta, onde δ_{ij} exprime a medida de dissemelhança entre os indivíduos i e j.

No caso das medidas de dissemelhança verificarem, além das três primeiras condições, a desigualdade triangular, está-se perante uma situação de medida de distância d_{ij}.

Uma medida de semelhança, s_{ij}, representando a dissemelhança entre os indivíduos i e j, caracteriza-se pelas seguintes propriedades:
1) $0 \leq s_{ij} \leq 1$: Determinado por $s_{ij} = 0$ quando os indivíduos não são semelhantes e por $s_{ij} = 1$ quando a semelhança é máxima.
2) $s_{ii} = 1$: A semelhança de um elemento a si próprio é zero.
3) $s_{ij} = s_{ji}$: A semelhança entre dois elementos é simétrica, onde s_{ij} exprime a medida de semelhança entre os indivíduos i e j.

Medidas de semelhança para variáveis categóricas

Para amostras em que os dados são, na sua maioria, categóricos, utilizam-se, com maior frequência, as medidas de semelhança. Estas medidas costumam fixar-se no intervalo $[0,1]$, sendo expressas, ocasionalmente, no intervalo de 0-100%. Dois indivíduos i e j têm coeficiente de semelhança $s_{ij} = 1$, se os dois tiverem valores idênticos para todas as variáveis. A semelhança 0 indica que os dois indivíduos diferem por completo em todas as variáveis. É possível fazer a transformação de medida de semelhança para dissemelhança a partir de $\delta_{ij} = 1 - s_{ij}$, por exemplo.

Medidas de semelhança para variáveis binárias

Estas medidas são indicadas para definir a semelhança entre os indivíduos de uma amostra multivariada, caracterizada por variáveis qualitativas, em especial, binárias. Considera-se que dois indivíduos são caracterizados por variáveis nominais dicotómicas, onde 1 e 0 significam, respetivamente, presença e ausência da característica. As medidas de semelhança entre os dois sujeitos i e j, baseiam-se, em geral, nas quatro quantidades seguintes que representam o número de variáveis, para os quais:
a - Ambos os sujeitos, i e j, tomam o valor 1;
b - O sujeito i toma o valor 1 e o sujeito j toma o valor 0;
c - O sujeito i toma o valor 0 e o sujeito j toma o valor 1;
d - Ambos os sujeitos, i e j, tomam o valor 0.

De entre os vários coeficientes de associação, destacam-se: coeficiente de concordância simples (*matching coefficient*), coeficiente de *Jaccard*, coeficiente de *Sørensen-Dice* e coeficiente de *Gower* e *Legendre*.

Tabela 2.3.2.1 - Tabela de contingência para variáveis binárias

	Indivíduo j			
	Resultado	1	0	Total
Indivíduo i	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

De acordo com a informação da Tabela 2.3.2.1, os coeficientes são definidos:

1) Coeficiente de concordância simples

$$s_{ij} = \frac{a+d}{(a+b+c+d)}\,, \qquad 0 \leq s_{ij} \leq 1 \tag{2.1}$$

Este coeficiente mede a semelhança entre dois indivíduos e é representado pela razão entre o número de características presentes e ausentes, simultaneamente, nos dois sujeitos e o número de características totais, situação que valoriza igualmente.

2) Coeficiente de *Jaccard*

$$s_{ij} = \frac{a}{a+b+c}\,, \qquad 0 \leq s_{ij} \leq 1 \tag{2.2}$$

Este coeficiente mede a semelhança entre dois indivíduos e não contempla o número de características neles ausentes.

3) Coeficiente de *Sørensen-Dice*

$$s_{ij} = \frac{2a}{(2a+b+c)}\,, \qquad 0 \leq s_{ij} \leq 1 \tag{2.3}$$

Este coeficiente mede a semelhança entre dois indivíduos, não contempla o número de características neles ausentes e dá o dobro do peso à variável em que ambos os sujeitos, i e j, tomam o valor 1.

4) Coeficiente de *Gower* e *Legendre*

$$s_{ij} = \frac{a}{\left[a+\frac{1}{2}(b+c)\right]}\,, \qquad 0 \leq s_{ij} \leq 1 \tag{2.4}$$

Mede a semelhança entre dois indivíduos, através da presença em ambos os sujeitos, não contempla o número de características ausentes, d, e retira importância às características, b e c, que dispõem de característica ausente ou presente ou vice-versa.

Medidas de dissemelhança/distância para variáveis contínuas

As medidas de dissemelhança propostas podem ser divididas em medidas de distância e medidas de correlação. De seguida, é apresentada a análise de alguns destes coeficientes, que são medidas de dissemelhança para dados contínuos usados nos algoritmos de agrupamento:

1) Distância de *Manhattan*[13]

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{2.5}$$

Representa-se por d_{ij} e traduz-se pela distância entre os elementos i e j, sendo p o número total de variáveis, x_{ik} o valor da variável k para o sujeito i $(k = 1, \dots, p)$ e x_{jk} o valor da variável k para o sujeito j $(k = 1, \dots, p)$.

2) Distância *Euclidiana*

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2} \tag{2.6}$$

Considera-se esta distância a medida de proximidade mais utilizada em situações de variáveis contínuas.

3) Quadrado da distância *Euclidiana*

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \tag{2.7}$$

Considera-se esta medida quando a distância entre dois casos, i e j, é definida como o somatório dos quadrados das diferenças entre os valores i e j para todas as variáveis.

4) Distância de *Chebyshev*

$$d_{ij} = \max_k |x_{ik} - x_{jk}| \tag{2.8}$$

A distância de *Chebyshev* é conhecida, também, por distância máxima do valor. Calcula a distância máxima da diferença entre as coordenadas de um par de indivíduos. Essa distância pode ser usada tanto para variáveis ordinais, como quantitativas.

[13] Designada também por distância absoluta ou *City-Block Metric*.

5) Distância de *Minkowski*

$$d_{ij} = \sum_{k=1}^{p} \left(\left| x_{ik} - x_{jk} \right|^{r} \right)^{\frac{1}{r}}, \qquad r \geq 1 \tag{2.9}$$

Tanto a distância Euclidiana ($r = 2$), como a de *Manhattan* ($r = 1$) são casos especiais desta distância. No caso em que $r \to \infty$, obtém-se a distância de *Chebyshev*.

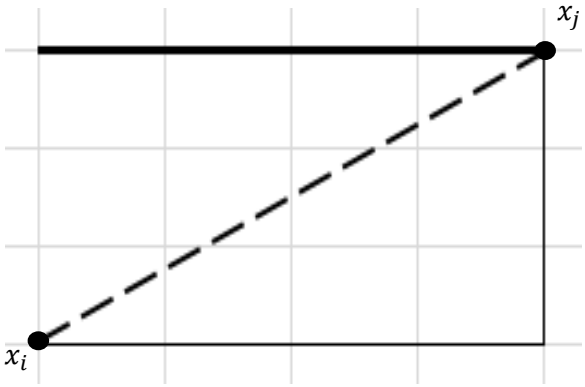

Figura 2.3.2.1 - Relação entre as distâncias *Euclidiana*, *Manhattan* e *Chebyschev*

A Figura 2.3.2.1 ilustra a relação entre as medidas de distância em análise. A distância *Euclidiana*, representada na linha a tracejado, é calculada através de uma linha reta entre os pontos x_i e x_j. A distância de *Manhattan* é calculada em quarteirão, unidade a unidade, e faz-se representar pela linha contínua fina. Por último, a distância de *Chebyschev*, linha contínua a negrito, dá-se pela maior das duas distâncias.

6) Separação Angular

$$\phi_{ij} = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\left(\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2 \right)^{\frac{1}{2}}}, \qquad -1 \leq \phi_{ij} \leq 1 \tag{2.10}$$

Em que 1 e -1 correspondem, respetivamente, a uma correlação perfeita positiva e negativa. Se o valor for 0, não existe relação.

7) Coeficiente de correlação de *Pearson*

$$\phi_{ij} = \frac{\sum_{k=1}^{p} (x_{ik} - \bar{x}_{ik}) \left(x_{jk} - \bar{x}_{jk} \right)}{\left[\sum_{k=1}^{p} (x_{ik} - \bar{x}_{ik})^2 \sum_{k=1}^{p} \left(x_{jk} - \bar{x}_{jk} \right)^2 \right]^{\frac{1}{2}}}, \qquad -1 \leq \phi_{ij} \leq 1 \tag{2.11}$$

Onde,

$\bar{x}_{ik} = \sum_{k=1}^{p} x_{ik} e^{-1} \leq \phi_{ij} \geq 1$, média de todas as variáveis para o sujeito i.
$\bar{x}_{jk} = \sum_{k=1}^{p} x_{jk} e^{-1} \leq \phi_{ij} \geq 1$, média de todas as variáveis para o sujeito j.

Nos casos em que $\phi_{ij} = 1$, o coeficiente indica a semelhança máxima, mas não necessariamente a identidade entre as características dos indivíduos i e j. Se $\phi_{ij} = -1$, indica o máximo de dissemelhança.

Os dois coeficientes apresentados, separação angular e coeficiente de correlação de *Pearson*, podem transformar-se em medidas de dissemelhança, através de:

$$\delta_{ij} = \frac{1 - \phi_{ij}}{2} \tag{2.12}$$

2.3.3. Métodos de formação de *clusters*

Depois de definidas as medidas de proximidade entre dois elementos, é necessário aplicar o mesmo procedimento às medidas de proximidade entre os *clusters*, ou seja, definir medidas de agregação entre os grupos (Everitt, Landau, Leese & Stahl, 2011).

A escolha do método de formação de *clusters* varia consoante o tipo de dados e o objetivo da análise. Existem dados onde é possível utilizar vários métodos de formação, não havendo argumentos suficientes para restringir a escolha do método mais adequado. Neste caso, aplicam-se os métodos candidatos e comparam-se os resultados obtidos, tendo também em conta a informação adicional existente sobre os dados. Este tipo de análise comparativa é muito importante, pelo facto de ser uma ferramenta descritiva e exploratória, que permite descobrir padrões interessantes nos dados e procurar informação relevante sobre a realidade e não provar ou refutar hipóteses preconcebidas (Kaufman & Rousseeuw, 1990).

Existem diversos métodos e algoritmos de *clustering*. Neste trabalho, apenas serão abordados os métodos hierárquicos e não hierárquicos, com especial foco nos métodos de partição, através da análise dos métodos *k-means* e *k-medoids*.

Métodos hierárquicos

Os métodos hierárquicos permitem a obtenção de *clusters* para indivíduos e variáveis, não sendo necessário definir o número de *clusters* inicial. Porém, se um elemento entrar num *cluster*, não poderá sair. Estes métodos dividem-se em dois tipos: aglomerativos e divisivos, sendo mais utilizados os métodos hierárquicos aglomerativos. Considera-se, desse modo, o processo aglomerativo quando se obtém, no final do mesmo, um único *cluster* com todos os elementos e divisivo quando, igualmente no final do processo, existirem n *clusters* com um único elemento cada um, como ilustrado na Figura 2.3.3.1.

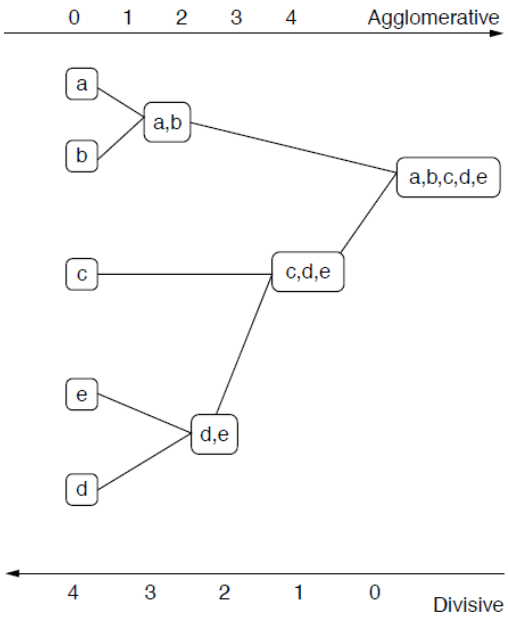

Figura 2.3.3.1 - Relação entre os métodos aglomerativo e divisivo (Kaufman & Rousseeuw, 1990)

Para qualquer um dos métodos, o objetivo é o mesmo: escolher a solução ótima, ou seja, o número ótimo de *clusters*, com a melhor separação possível.

O resultado de um processo hierárquico pode ser representado através de um diagrama bidimensional – Dendrograma (Hand *et al.*, 2001).

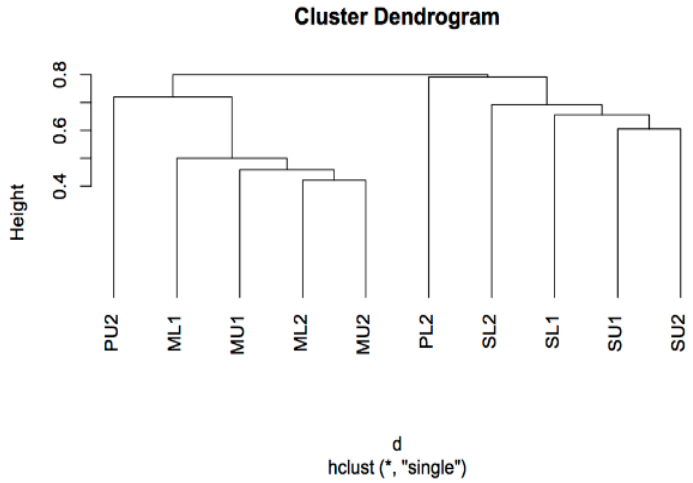

Figura 2.3.3.2 - Exemplo de um dendrograma

Cada ramo do dendrograma representa um elemento e a raiz a aglomeração de elementos. É possível, através da representação gráfica do processo de criação de *clusters*, presente na Figura 2.3.3.2, identificar os *clusters* agrupados ao longo de todo o processo – trajeto vertical – e observar o incremento nos valores da distância entre os *clusters* – trajeto horizontal.

Os métodos aglomerativos iniciam-se com cada elemento no seu próprio *cluster* e a cada iteração vão-se aglomerando esses mesmos elementos. O método acaba, quando existir uma aglomeração de todos

os elementos num só *cluster* ou estiver cumprida alguma condição de paragem pré-estabelecida. Essa aglomeração progressiva é feita através de uma medida de similaridade.

Alguns exemplos de métodos aglomerativos estão ilustrados na Figura 2.3.3.3. Cada representação corresponde às definições apresentadas de seguida:

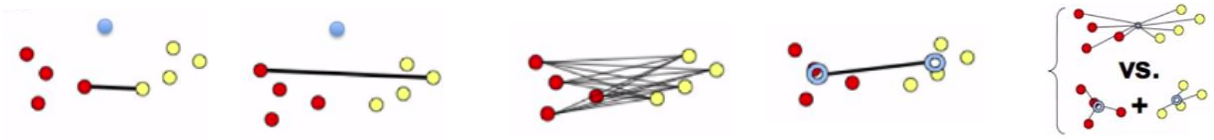

Figura 2.3.3.3 - Medidas de proximidade entre *clusters* (Lavrenko, 2015)

1) *Single Linkage*
A distância entre dois grupos é medida através do mínimo entre um par de indivíduos, x e y, entre dois os *clusters*, C_i e C_j, onde $x \in C_i, y \in C_j$. A utilização deste critério tende a resultar em *clusters* desequilibrados e desalinhados, especialmente no caso em que existe um elevado número de dados.

$$d_{ij} = \min_{x \in C_i, y \in C_j,} d(x,y) \tag{2.13}$$

2) *Complete Linkage*
A distância entre dois grupos é medida como sendo a distância máxima entre um par de indivíduos, entre todos os *clusters*. Utilizando este critério, são formados *clusters* mais compactos.

$$d_{ij} = \max_{x \in C_i, y \in C_j,} d(x,y) \tag{2.14}$$

3) *Average Linkage*
A distância entre dois grupos traduz-se pela média da distância entre todos os pares de indivíduos dos dois grupos. Através deste critério, os *clusters* obtidos terão pequenas variações de distância.
Este é um critério intermédio, entre o *Complete Linkage* e o *Single Linkage*.

$$d_{ij} = \operatorname{sum}_{x \in C_i, y \in C_j,} \frac{d(x,y)}{n_i \times n_j} \tag{2.15}$$

4) Critério do Centróide
Esta distância é medida através da distância entre os seus centróides, sendo que cada centróide corresponde ao ponto médio no espaço definido pelos grupos.

$$d_{ij} = \left\| \bar{x}_i - \bar{y}_j \right\|^2 \tag{2.16}$$

5) Critério de *Ward*

Neste caso, não são calculadas distâncias. Este método assume que um *cluster* é representado pelo seu centróide, r, e procura aglomerar os dois *clusters*, C_i e C_j, que minimizam a soma das distâncias quadráticas dos indivíduos do *cluster*, aos seus centróides.

$$d_{ij} = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} \left(x - r_j\right)^2 + \sum_{x \in C_{ij}} \left(x - r_{ij}\right)^2 \tag{2.17}$$

O agrupamento aglomerativo é dado pela matriz de proximidades $D = \left[d_{ij}\right]$, de ordem $n \times n$. Seguem-se, segundo Timm (2002), quatro passos para a sua elaboração:

1. Iniciar o processo com n *clusters*, cada um deles com um elemento;
2. Usar uma matriz D, escolhendo os elementos mais semelhantes, i e j;
3. Através da junção desses dois elementos, i e j, é formado um novo *cluster* (ij). Recalcular as distâncias entre o novo *cluster* (ij) e os elementos já existentes, usando o critério de agregação escolhido. Consequentemente, é obtida uma nova matriz de proximidade de ordem $(n-1) \times (n-1)$.
4. Repetir os passos 2 e 3, $(n-1)$ vezes.

O método divisivo é iniciado com a existência de um único *cluster* e a cada iteração vai sendo dividido em *clusters* mais pequenos, até existir uma paragem ou até que o número de *clusters* seja igual ao número de elementos existentes. Este método funciona, geralmente, de forma oposta aos métodos aglomerativos e é utilizado, normalmente, quando o objetivo visa partições de *clusters* extensos.

Numa fase inicial, é importante ter em consideração todas as possíveis divisões dos dados em dois grupos. Porém, essa pode ser uma solução inviável, uma vez que o número de combinações será extremamente elevado. Por isso, este método é pouco utilizado.

Os métodos hierárquicos proporcionam uma descrição sequencial de como os indivíduos se relacionam entre si. Contudo, são portadores de uma grande incapacidade para alterar uma má escolha. Caso os *clusters* sejam afetados por decisões menos boas, tomadas nos primeiros passos, é impossível recuar e corrigi-la.

Métodos não hierárquicos

Os métodos não hierárquicos são utilizados na construção de *clusters* de indivíduos. Nestes métodos, agrupam-se os n elementos em k *clusters* e a cada iteração é feita uma melhoria ou mudança dos elementos entre *clusters*, sendo necessário escolher o número de *clusters* no início do processo, mantendo, assim, esse número fixo. O processo finaliza quando cada elemento estiver no *cluster* mais apropriado, em relação à técnica usada.

Existem diferentes métodos não hierárquicos, com o seu desempenho a depender não só da primeira agregação dos indivíduos em *clusters*, mas também da forma como as novas distâncias entre os centróides dos *clusters* e indivíduos é calculada.

Estes métodos subdividem-se em quatro: método de partição, método baseado em modelos, método difuso e método de sobreposição. A grande diferença entre eles está na forma como se desenrola a primeira agregação dos indivíduos em *clusters* e na maneira como as distâncias entre os centróides dos *clusters* e os indivíduos são realizadas.

Método de partição

Este método aplica-se, essencialmente, a indivíduos, opera através de uma matriz de dados e exige, tal como nos restantes métodos não hierárquicos, que o número de grupos seja fixado. Não utiliza a matriz de proximidades inicial, apenas a matriz de dados inicial – o oposto aos métodos hierárquicos.

A partir de um conjunto de dados é definida uma partição em que os *clusters* têm obrigatoriamente de satisfazer os critérios de homogeneidade, coesão interna, isolamento dos grupos e heterogeneidade.

Estes métodos são muitas vezes utilizados como complemento aos métodos hierárquicos, com o objetivo de ajudar a escolher o corte ótimo na estrutura de dados, de forma a conseguir dar resposta à questão: "Quantos *clusters* apresentam?".

Um dos métodos partitivos mais frequente nos *softwares* estatísticos é o *k-means*, que mede a proximidade entre grupos usando a distância *Euclidiana* entre os centróides dos grupos. É apresentado nos seguintes passos (Johnson & Wichern, 2002):
1. Dividir os objetos em *k clusters* iniciais e calcular os centróides;
2. Atribuir cada objeto ao *cluster* cujo centróide está mais próximo;
3. Recalcular o centro de cada grupo;
4. Repetir os passos 2 e 3 até nenhum elemento mudar de grupo.

Esta técnica é bastante utilizada, pois consegue operar em conjuntos de dados de grande dimensão e, normalmente, converge rapidamente. A eventual desvantagem resulta da procura única de *clusters* esféricos de dimensão igual.

Outro dos métodos mais utilizados neste tipo de análises é o *k-medoids*, através do algoritmo *PAM*. Este método, tal como o *k-means*, centra-se na partição dos dados, permitindo a obtenção de *clusters*, após definido o número *k* de *clusters* pretendidos. Este método visa a procura de indivíduos representativos dos *clusters*, medóides, de entre todos os indivíduos do conjunto de dados.

Após o cálculo dos medóides, formam-se os conjuntos de dados, atribuindo cada indivíduo do *cluster* ao medóide mais próximo (Kaufman & Rousseeuw, 1990). Os medóides, por sua vez, são indivíduos representativos de um conjunto de dados ou de um *cluster*, com um conjunto de dados cuja dissimilaridade média, a todos os indivíduos no *cluster*, é mínima. São normalmente utilizados em dados onde não é possível definir o centróide e fazem sempre parte do conjunto de dados.

O algoritmo pode ser apresentado através dos seguintes passos:
1. Dividir os indivíduos em *k clusters* iniciais e calcular os seus medóides;
2. Atribuir cada indivíduo aos medóides mais próximo;
3. Calcular os medóides, de cada grupo, com a menor distância;
4. Repetir os passos 2 e 3 até nenhum elemento mudar de grupo.

Este algoritmo tem como objetivo minimizar a média das dissemelhanças entre os indivíduos e os seus medóides mais próximos.

Para encontrar os medóides, são efetuados os seguintes passos (Kaufman & Rousseeuw, 1990):
1. Considerar um indivíduo x_i, ainda não selecionado, como candidato a medóide inicial;
2. Para um indivíduo x_j, não selecionado, calcular a dissemelhança, d_j, entre o indivíduos x_j e o indivíduo mais próximo selecionado previamente;
3. Se $d_j = d(x_i, x_j)$, então o indivíduo x_j vai contribuir para a decisão de seleção do indivíduo x_i, já que melhora o conjunto dos medóides iniciais. Seja $C_{ij} = \max\{d_j - d(x_j, x_i), 0\}$;
4. Calcular o ganho total, $\sum_j C_{ij}$, obtido, selecionando o indivíduo x_i;
5. Escolher o indivíduo x_i, que maximize $\sum_j C_{ij}$.
Este processo continua até serem encontrados os k medóides.

Estes dois modelos, apresentados em cima, pertencentes aos métodos de partição, atribuem cada indivíduo a um único *cluster* e estabelecem conjuntos de dados baseados em protótipos, ou seja, centróides para as *k-means* e medóides para os *k-medoids*.

A grande vantagem destes modelos é a possibilidade de poderem ser aplicados a conjuntos de grande dimensão de dados. Porém, apresentam a desvantagem de não conseguirem suportar *clusters* não esféricos e com dimensões diferentes. Para além disso, não apresentam bons resultados quando o conjunto de dados apresentado é compreendido por indivíduos bastante dissimilares (Wei, Lee & Hsu, 2003).

Medidas de coesão e separação

As medidas de coesão e separação, incorporadas nos métodos de formação de *clusters*, pertencem às medidas de validação para partições de dados, baseadas em protótipos, nos casos em que os *clusters* podem ser representados pelos seus centróides ou medóides. Estas medidas de validação podem ser aplicadas a um conjunto de dados, mas também na validação individual de *clusters* ou indivíduos. Deste modo, são avaliados os indivíduos pertencentes a um *cluster*, para apurar qual o seu contributo para a coesão ou separação desse *cluster*.

Existem vários índices que podem ser aplicados, com o objetivo de se obter o corte ótimo ao conjunto de dados. Estas medidas permitem avaliar indivíduos, *clusters* e conjuntos de *clusters*.

O *package NbClust*, da autoria de Charrad, Ghazzali, Boiteau e Niknafs (2014), oferece 30 índices que permitem determinar o número de *clusters* ideal e propõe ao utilizador o melhor esquema de agrupamento dos diferentes resultados obtidos, variando todas as combinações de número de *clusters*, medidas de distância e métodos de agrupamento.

De seguida, é apresentado um pequeno resumo sobre os índices mais relevantes para a análise de dados, que apenas podem ser aplicados a dados em que foram utilizados métodos hierárquicos. Torna-se, por isso, necessário saber identificar o mais adequado, entendendo os diferentes modos de funcionamento de cada um deles, para a correta aplicação na componente prática deste trabalho.

Índice de *Frey*

Este índice, desenvolvido por Frey e Van Groenewoud, em 1972, apenas pode ser aplicado em métodos hierárquicos. De acordo com a função representada em baixo, é calculado através da diferença entre dois níveis sucessivos na hierarquia, em que o numerador representa a diferença entre as distâncias médias *intercluster*, $\bar{d}_b$, entre dois níveis da hierarquia (nível j e $j+1$). O denominador caracteriza-se pela diferença entre a média das distâncias dentro do *cluster*, $\bar{d}_w$, entre dois níveis (nível j e $j+1$). Os autores descrevem o rácio de 1,00, para identificar o corte correto, enquanto as proporções podem variar acima e abaixo de 1,00. Se o rácio nunca tingir 1,00, então a solução de *cluster* único é assumida (Milligan & Cooper, 1985).

$$Frey = \frac{\bar{S}_{bj+1} - \bar{S}_{bj}}{\bar{S}_{wj+1} - \bar{S}_{wj}} \tag{2.18}$$

$\bar{S}_b = \frac{S_b}{N_b}$ Distância média *interclusters.*

$\bar{S}_w = \frac{S_w}{N_w}$ Distância média *intracluster.*

Índice de *McClain*

Criado por McClain e Rao, em 1975, consiste na razão de dois termos. O numerador corresponde à média da distância no *cluster*, dividida pelo número de distâncias no *cluster*. O denominador é a média entre a distância do *cluster*, dividida pelo número de distâncias no *cluster*.

$$McClain = \frac{\bar{S}_w}{\bar{S}_b} = \frac{S_w/N_w}{S_b/N_b} \tag{2.19}$$

O valor mínimo do índice é o indicado para o número ótimo de *clusters*.

Índice de *Cindex*

O *C-Index* foi revisto por Hubert e Levin, em 1976. É calculado usando a razão entre a soma das distâncias *interclusters*, S_w, divida pela soma das menores distâncias entre todos os pares de pontos no conjunto de dados, S_{min}, e a divisão entre a soma das maiores distâncias, S_{max}, entre todos os pares de pontos, em todo o conjunto de dados e S_{min}.

$$Cindex = \frac{S_w/S_{min}}{S_{max}/S_{min}}, S_{min} \neq S_{max}, Cindex \in (0,1) \tag{2.20}$$

O valor mínimo do índice é usado para indicar o número ótimo de *clusters* (Milligan & Cooper, 1985).

Índice de *Silhouette*

Este índice foi introduzido por Rousseeuw, em 1987. O índice de *Silhouette*, para determinado *cluster*, é definido pela média dos coeficientes dos indivíduos desse *cluster*.

$$Silhouette = \frac{\sum_{i=1}^{n} S(i)}{n}, Silhouette \, \epsilon \, [-1,1] \tag{2.21}$$

Onde,

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}$$

$a(i) = \frac{\sum_{j \, \epsilon \, \{Cr/i\}} dij}{n_r - 1}$, é a dissimilaridade média do i indivíduo a todos os outros indivíduos do *cluster* Cr.

$b(i) = \min_{s \neq r} \{d_{iC_s}\}$

$d_{iC_s} = \frac{\sum_{j \, \epsilon \{C_s\}} dij}{n_s}$, é a dissimilaridade média do *i*-ésimo indivíduo para todos os indivíduos do *cluster* Cs.

O valor do índice pode variar entre -1 e 1, onde os valores negativos correspondem à situação em que $a(i)$, a dissimilaridade média dos indivíduos do *cluster*, é maior do que $b(i)$, a distância média mínima a indivíduos de outro *cluster*. O objetivo é que este índice seja positivo, ou seja, $a(i) < b(i)$, e que o valor de $a(i)$ seja tão próximo de 0 quanto possível, pois o coeficiente atinge o seu valor máximo quando $a(i) = 0$. O valor máximo do índice é usado para determinar o número ótimo de *clusters* nos dados. $S(i)$ não é definido para $k = 1$, caso em que admite apenas um *cluster*, não havendo assim nenhum corte ao conjunto de dados (Kaufman & Rousseeuw, 1990).

Índice de D*unn*

O índice de *Dunn*, criado por J. C. Dunn, em 1974, é representado pelo rácio entre a distância *intercluster* mínima e a distância *intracluster* máxima.

$$Dunn = \frac{\min\limits_{1 \leq i \leq j \leq q} d(C_i, C_j)}{\max\limits_{1 \leq k \leq q} diam(C_k)}, \tag{2.22}$$

Onde $d\,(C_i, C_j)$ é a função de dissimilaridade entre dois *clusters*, C_i e C_j, definidos como $d(C_i, C_j) = \min\limits_{x \, \epsilon \, C_i, y \, \epsilon \, C_j} d(x, y)$ e $diam(C)$ é o diâmetro de um *cluster*, que pode ser considerada uma medida de dispersão de *cluster*. O diâmetro de um *cluster* C pode ser definido da seguinte forma: $diam(C) = \max\limits_{x, y \, \epsilon \, C} d(x, y)$.

Assim, se o conjunto de dados tiver *clusters* compactos e bem separados, o diâmetro dos *clusters* deverá ser pequeno e a distância entre os agrupamentos grande. Desta forma, o índice de *Dunn* deve ser maximizado, para que seja possível obter o número ótimo de *clusters*, de modo a encontrar uma partição adequada dos dados – note-se que o resultado obtido compreende valores entre 0 e infinito. A sua principal desvantagem é ser um índice de elevada complexidade[14] computacional.

[14] A complexidade varia consoante o número de *k clusters* e o número de variáveis utilizadas para caracterizar os indivíduos.

2.4. *Market Basket Analysis*

Market basket analysis designa-se, em português, por análise ao cabaz de compras. Esta técnica de *data mining* utiliza regras de associação para identificar os hábitos de compra dos consumidores. Para isso, analisa os artigos com maior frequência de compra, mas também relações de complementaridade e alternativas de substituição, por forma a aplicar essas conclusões em técnicas de marketing, com destaque para a venda cruzada.

Existem vários setores de negócio, onde é possível colocar em prática esta técnica. Por exemplo, nos corredores de supermercado, a colocação estratégica de *toppings* e rolos de bolacha junto aos gelados, sugere a compra conjunta dos artigos. Também alguns mecanismos de compra *online*, através de *sites* especializados, apresentam recomendações de artigos, com base nos históricos de pesquisa e compra.

Para se perceber como são realizadas este tipo de análises, é necessário observar, em detalhe, de que forma as regras de associação são aplicadas a um conjunto de dados. Nesse sentido, analisa-se um conjunto de produtos alimentares, apresentado através do seguinte exemplo:

Seja $X = \{x_1, x_2, \dots, x_p\}$ um conjunto de p artigos disponíveis para serem comprados, em simultâneo, em cada transação t. O conjunto de todas as transações, $T = \{t_1, \dots, t_n\}$, onde n é o número de transações composto por listas de artigos, todos identificados pelo *ID* associado. Cada transação corresponde a um dado momento, em que um consumidor compra um ou vários produtos disponíveis. Um *itemset* corresponde ao conjunto de artigos comprados em cada transação.

Tabela 2.4.1 - Resumo das transações efetuadas por cinco consumidores

Transações	Produtos
t_1	$\{leite, iogurte, manteiga, queijo\}$
t_2	$\{iogurte, queijo\}$
t_3	$\{leite, bolachas, manteiga\}$
t_4	$\{leite, iogurte, manteiga\}$
t_5	$\{iogurte\}$

Na Tabela 2.4.1, é apresentado um exemplo de cinco transações, efetuadas por cinco consumidores, com cinco produtos distintos. Apesar de ser um exemplo fictício, aplica-se de igual forma o procedimento habitual, como se de uma análise a dados reais se tratasse. Dessa forma, é necessário fazer uma triagem das transações que se pretende estudar.

As regras de associação iniciam-se pela regra $\{X \rightarrow Y\}$, em que X é definido como o antecedente e Y o consequente. A regra representa a relação de compra entre o *itemset* X e o *itemset* Y e indica a probabilidade de Y ser adquirido quando X foi comprado (Gama *et al.*, 2012). Por exemplo, $leite \rightarrow queijo$, se o consumidor compra leite, então, também compra queijo.

Posto isto, torna-se necessário explicar alguns dos conceitos utilizados para a realização da *market basket analysis*. Assim, define-se como *support* a frequência com que os artigos, numa dada regra, ocorrem em simultâneo. É calculado da seguinte forma:

$$Support(X,Y) = \frac{frq(X,Y)}{n} \tag{2.23}$$

O conceito de *support* divide-se em dois tipos em que X é o antecedente e Y o consequente. O *absolute support*, correspondente à frequência absoluta, representa o número de transações onde um dado artigo aparece no *itemset* e o *relative support*, referente a proporção de transações que contem esse mesmo *itemset*. O *relative support* é calculado através da divisão entre o *absolute support* e o número total de transações (Gama *et al*., 2012).

A *market basket analysis* realiza-se em duas fases. A primeira tem como objetivo encontrar conjuntos frequentes de artigos comprados, definindo o *support* mínimo, sendo que prevalecem aqueles que tiverem *support* igual ou superior ao mínimo estabelecido. Na segunda, são definidas as regras de associação, consoante o valor *confidence* fixado (Ulas, 1999).

Assim, observando a Tabela 2.4.1, verifica-se 3 o *absolute support* do leite, sendo o seu *relative support* calculado da seguinte forma $^{3}/_{5} = 0{,}6$, sendo este o valor das transações efetuadas que contêm leite.

Como forma de encontrar as transações mais relevantes para a realização da análise, fixa-se um número mínimo para o *absolute support*. Se o número mínimo for fixado em 2, então as transações com relevância, *itemsets* frequentes, são apresentadas como demonstra a Tabela 2.4.2.

Tabela 2.4.2 - *Itemsets* frequentes com *support* mínimo de 2

1 *item*	2 *items*	3 *items*
$\{Leite\}$: 3	$\{Leite, Iogurte\}$: 2	$\{Leite, Iogurte, Manteiga\}$: 2
$\{Iogurte\}$: 4	$\{Leite, Manteiga\}$: 3	
$\{Manteiga\}$: 2	$\{Iogurte, Manteiga\}$: 2	
$\{Queijo\}$: 3	$\{Iogurte, Queijo\}$: 2	

Observando a Tabela 2.4.2, que apresenta os *itemsets* frequentes, quando se define o *support* mínimo em 2, verifica-se que foram obtidos nove conjuntos frequentes de artigos, para as diferentes dimensões apresentadas. O produto Bolachas foi excluído dos conjuntos frequentes, por não ter *support* mínimo, visto que apenas aparece numa transação.

Desse modo, e depois de encontrados os conjuntos frequentes de dados, procede-se à derivação de regras de associação. Nesta segunda fase, calculam-se as regras que expressam a coocorrência provável de produtos, dentro de conjuntos de artigos frequentes. Através do algoritmo *Apriori*, calcula-se a probabilidade de um artigo estar presente num *itemset* frequente.

A métrica *confidence* indica a probabilidade do antecedente e do consequente ocorrerem na mesma transação e calcula-se através da seguinte fórmula:

$$Confidence\ (X \rightarrow Y) = \frac{supp(X,Y)}{supp(X)} \tag{2.24}$$

Aplica-se esta fórmula ao exemplo que tem vindo a ser apresentado, calculando-se a *confidence*, grau de confiança para os *itemsets* frequentes, a 2 *items*. Sendo representado pela proporção de vezes em que

o consumidor compra X, também, compra Y.

Tabela 2.4.3 - Algumas regras de associação realizadas a partir dos *itemsets* frequentes

Regra	*Confidence*
$\{Leite \rightarrow Iogurte\}$	$2/3 = 0{,}67$
$\{Leite \rightarrow Manteiga\}$	$3/3 = 1$
$\{Iogurte \rightarrow Queijo\}$	$2/4 = 0{,}5$

Através da Tabela 2.4.3, é possível verificar, observando que o grau de confiança é 1, para a regra $Leite \rightarrow Manteiga$, em que todas as transações que incluem Leite, também contêm Manteiga, ou seja, os consumidores que se deslocam ao supermercado, por exemplo, para comprar Leite, compram sempre Manteiga. Na regra $\{Iogurte \rightarrow Queijo\}$, conclui-se que em metade das deslocações que o consumidor faz ao supermercado, por exemplo, para comprar Iogurte, acaba também por adquirir Queijo. É necessário ter-se em conta a rotatividade do artigo em questão e interpretar os resultados obtidos.

Mesmo após definidos os valores mínimos para *support* e *confiance*, de forma encontrar transações relevantes, isso não significa que as regras de associação encontradas sejam em número reduzido. Deste modo, caso as regras obtidas sejam demasiado extensas, utilizam-se medidas adicionais, que permitem validar a qualidade da regra estudada, tal como o *lift* (Hahsler & Hornik, 2007).

O cálculo do *lift* é feito através do quociente entre o *support* de regra de associação e o *support* do antecedente e consequente. Esta métrica mede a diferença do número de vezes que o antecedente e o consequente são adquiridos em simultâneo.

Esta métrica tem em conta a dependência entre as variáveis em estudo, considerando que se duas variáveis são independentes, então a regra inferida não tem interesse. Isto acontece porque a associação entre os dois *itemsets* é calculada de forma aleatória, utilizando a seguinte fórmula, em que o grau de dependência varia no intervalo $[0, +\infty]$:

$$Lift\ (X \rightarrow Y) = \frac{Support\ (X, Y)}{Supp(X) \times Supp(Y)} \tag{2.25}$$

Assim sendo, se o resultado for inferior a 1, significa que os produtos analisados estão negativamente relacionados. Por outro lado, valores superiores a 1 indicam que os produtos estão positivamente relacionados. Se os valores estiverem próximos de 1, está-se perante produtos com fraca relação entre si, sendo, por isso, independentes (Hahsler & Hornik, 2007).

Após definidas as métricas para aplicar aos dados – *support*, *confidence* e *lift* –, com o objetivo de analisar as regras de associação, na maioria dos casos, o conjunto de dados é bastante extenso e, por isso, é necessária a utilização de algoritmos para a análise.

Assim, utiliza-se o algoritmo *Apriori*, que, para além de ser uma das ferramentas pioneiras, é também uma das mais rápidas de implementar (Ulas, 1999). Este algoritmo começa por criar *itemsets* com *support* superior ao definido por cada nível. Após definidos os candidatos a *itemsets* frequentes, o algoritmo testa a sua frequência, percorrendo a base de dados da transação.

Depois, o algoritmo cria as regras de associação, derivadas dos *itemsets* frequentes, considerando o valor mínimo definido para a *confidence*. Este algoritmo pode ser otimizado uma vez que, quando o *itemset* é movido do antecedente para o consequente, a *confidence* não pode aumentar. Assim, analisam-se apenas os descendentes de *itemsets* frequentes, visto que, se um *itemset* não for frequente, ainda menos serão os seus descendentes (Rodrigues, Gama & Ferreira, 2012).

As regras de associação são a forma mais tradicional de estudar grandes conjuntos de dados. Porém, este tipo de métodos produz regras em excesso (Zaki, 2000), que, por vezes, não são relevantes, aumentando o tempo de computação e a dificuldade em retirar conclusões.

Dessa forma, utiliza-se o *software* estatístico *Rstudio*, para a análise de regras de associação, através do *package arules*. Este pacote permite encontrar conjuntos de dados frequentes e produzir regras de associação, através de diferentes algoritmos, entre eles o algoritmo *Apriori* (Hahsler, Grun & Hornik, 2005), mencionado anteriormente. Este *package* será aplicado, e devidamente aprofundado, no Capítulo 6 deste trabalho, referente à análise de regras de associação.

Capítulo 3

Análise dos dados

3.1. Análise exploratória

De forma a poder ter-se uma melhor perspetiva dos dados a analisar, torna-se necessário explorar todas as componentes do *raw data*, tendo como objetivo o início do contacto com os dados. Nesse sentido, esta análise terá duas vertentes: caracterização através do perfil de lares existentes em Portugal Continental e respetivas compras realizadas.

3.1.1. Caracterização dos lares

Para a caracterização dos 447 lares, teve-se em conta as cinco características sociodemográficas que estão associadas a cada família: área de residência, número de membros, presença de crianças, classe social e tipo de família.

O painel de lares está dividido pelas diferentes regiões *Nielsen*, em Portugal Continental, com 30% da amostra a residir na área 1, 21% na área 3N e 13% na área 2. Assim, mais de 50% dos lares compreende-se nas regiões da Grande Lisboa, do Porto e no Litoral Norte do país, como demonstra o Gráfico 3.1.1.1. As restantes famílias, 36%, encontram-se no Interior, Litoral Sul e Algarve.

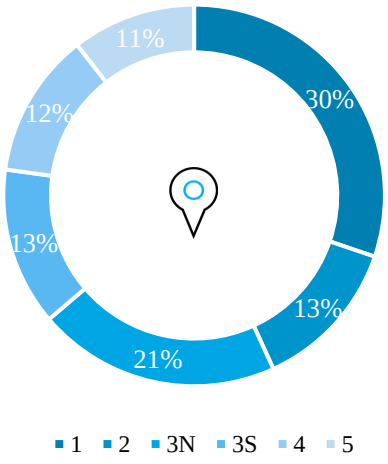

Gráfico 3.1.1.1 - Área de residência

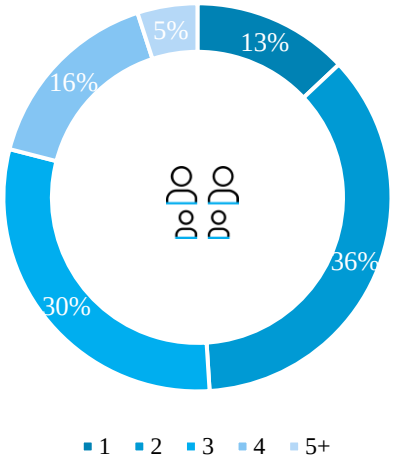

Gráfico 3.1.1.2 - Número de membros

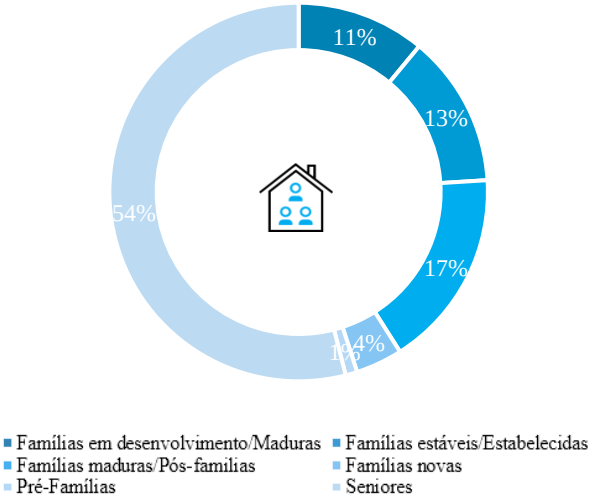

Gráfico 3.1.1.3 - Tipo de família

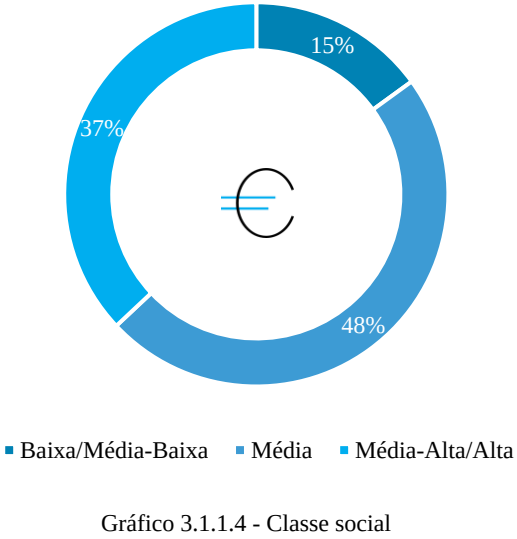

Gráfico 3.1.1.4 - Classe social

Há 50 anos, a dimensão média das famílias portuguesas atingia 3,7 membros. De acordo com o relatório *Censos 2011*, o último do género realizado, em Portugal, esse valor diminuiu para 2,6 pessoas (Instituto Nacional de Estatística [INE] & Pordata, 2015).

Através da análise do Gráfico 3.1.1.2, é possível verificar que a maioria, mais de 60%, da amostra vai ao encontro da dimensão média real das famílias portuguesas, sendo constituída por 2 ou 3 membros por lar, representados por 36% e 30%, respetivamente. Existe ainda uma outra fração relevante, correspondente a 21%, que compreende 4 ou mais membros na família.

Quanto à estrutura etária portuguesa, verificou-se, entre 2001 e 2011, um aumento da população idosa com 65 ou mais anos (INE, 2011). De acordo com o Gráfico 3.1.1.3, os Seniores são, justamente, o tipo de família predominante nesta amostra, com 52%, seguindo-se as Famílias maduras/Pós-famílias, 17%, e as Famílias estáveis/Estabelecidas, 13%.

Por outro lado, o número de lares com presença de crianças é bastante baixo. De acordo com dados recentes do INE (2017), referentes ao período em análise, a percentagem de famílias com crianças é de aproximadamente 46% enquanto que na amostra esta percentagem é de apenas 23%. Esta variável acaba, assim, por sofrer influência direta do facto de mais de metade da amostra se situar no tipo de família Seniores.

Em Portugal, a classe social Média tem vindo a perder dimensão. De acordo com dados da *Marktest* (2010), a classe baixa é predominante, seguida da classe média-baixa. Relativamente à amostra utilizada, encontram-se apenas três escalões de análise: Baixa/Média-Baixa, Média e Média-Alta/Alta. A classe social que prevalece é a Média, com 48%, contrariamente ao que acontece, em Portugal (Marktest, 2010), seguindo-se a Baixa/Média-Baixa, com 37%, e a Média-Alta/Alta, com 15% do total da amostra.

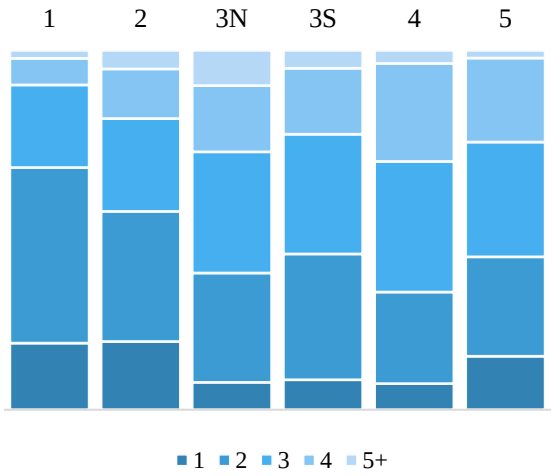

Gráfico 3.1.1.5 - Área de residência/Número de membros

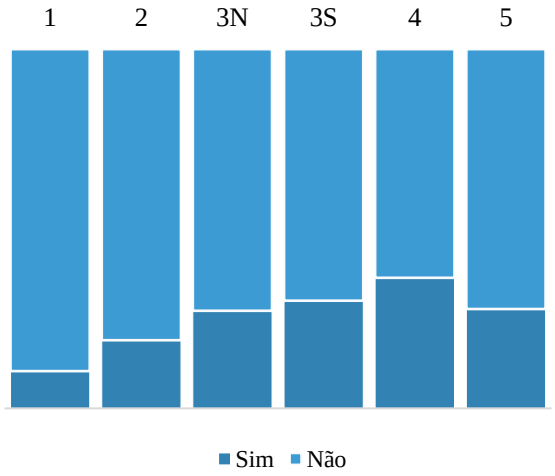

Gráfico 3.1.1.6 - Área de residência/Crianças

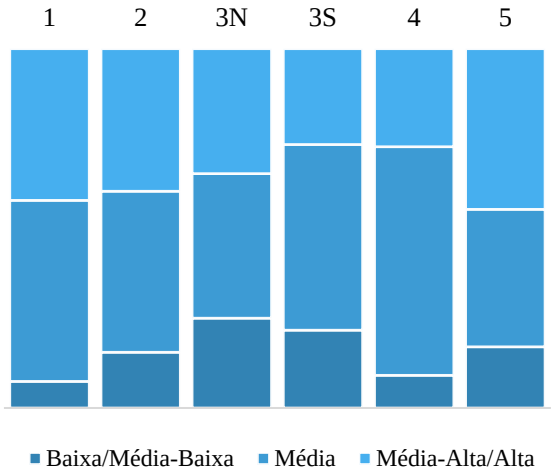

Gráfico 3.1.1.7 - Área de residência/Crianças

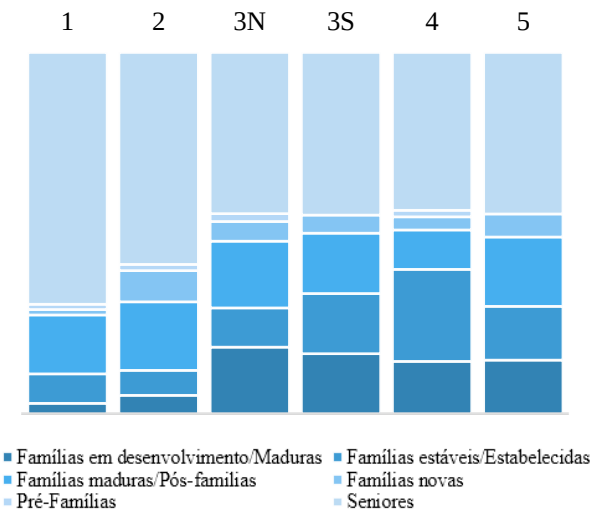

Gráfico 3.1.1.8 - Área de residência/Tipo de família

Realizou-se uma análise bivariada aos dados, entre a área de residência e as restantes quatro características sociodemográficas, por forma a identificar padrões entre a relação das características com a variável Área de residência.

Através da observação dos gráficos apresentados acima, é possível verificar que não existem semelhanças entre as seis áreas de residência e as quatro características em estudo. Conclui-se, portanto, que as regiões analisadas são heterogéneas, comportam-se de maneira diferente e são constituídas por famílias diferentes, não existindo, por isso, duas regiões idênticas.

A região 1 representa a área metropolitana da Grande Lisboa e é consituída, maioritariamente, por famílias pequenas de 2 elementos e sem crianças. A classe social predominante insere-se em ambas as classes Média e Média Alta/Alta, sendo esta uma região com elevado poder económico. Em relação ao tipo de família, esta mostra-se envelhecida, representando a região com mais Seniores entre todas as regiões *Nielsen*.

A área metropolitana do Grande Porto, representada pela região 2, é caracterizada por famílias pequenas, maioritariamente com 2 e 3 membros, e com poucas crianças. Apresenta uma classe social semelhante à região 1, constituída, porém, por mais famílias de classe Baixa/Média-Baixa. Por outro lado, o tipo de família difere da presente na área metropolitana da Grande Lisboa, ao incluir Famílias novas na fração, tornando esta zona a que maior número de famílias deste tipo compreende. Tem, ainda, e tal como todas as outras regiões, uma proporção elevada de famílias Seniores.

A região 3N, área Litoral Norte do país, apresenta-se como a mais numerosa, pois é a que mais famílias com 4 ou mais elementos inclui. É também a região que menos famílias com apenas um membro tem. Nesse sentido, quando comparada com as regiões 1 e 2, por exemplo, tem uma maior proporção de crianças. É considerada a zona do país com menor poder de compra, sendo a região onde a proporção da classe social Baixa/Média Baixa é maior, predominando, ainda assim, a classe Média. A região 3N é a que tem mais Famílias em desenvolvimento/Maduras, tendo, contudo, os Seniores a maior representação, no que ao tipo de família diz respeito.

Na área Litoral Sul, região 3S, o comportamento do número de membros por lar é semelhante à área 3N, onde predominam as famílias com 2 membros e com poucas crianças. A classe social predominante é a Média. O tipo de família com maior representatividade são, uma vez mais, os Seniores, enquanto, por outro lado, as Pré-famílias nesta zona são inexistentes. As Famílias maduras/Pós-famílias, Famílias estáveis/Estabelecidas e Famílias em desenvolvimento/Maduras têm a mesma representatividade nesta região.

O Interior Norte do país, zona 4, é constituído por poucas famílias com 1 ou 2 elementos, predominando os lares com 3 membros. Para além disso, é a área em análise onde a presença de crianças é maior. Em relação à classe social, a Média Alta/Alta é a menos presente de entre as três classes. Por outro lado, o Interior Norte revela-se a zona onde a fração da classe Média, entre todas as áreas, é maior. Esta zona abrange, ainda, mais Famílias estáveis/Estabelecidas do que as restantes, detendo, porém, o menor número de Famílias maduras/Pós-famílias.

Por último, no Sul de Portugal, representado pela região 5, predominam as famílias com 3 elementos. No entanto, lares com 5 ou mais membros quase não estão aqui representados, sendo esta a região que menos detém este tipo de famílias. Quanto à presença de crianças, apresenta uma proporção razoável, quando comparada com outras regiões. A classe prevalecente é a Média-Alta/Alta. As Pré-famílias não

habitam nesta zona, contrariamente às famílias Seniores, que são o tipo de família predominante. Apesar disso, esta é a área em que a proporção de famílias idosas é menor, quando comparada com as restantes.

3.1.2. Caracterização das compras realizadas

A análise exploratória não fica completa sem a análise ao tipo de consumidor, ou seja, a todos os seus comportamentos relativamente às compras realizadas. Foi necessário explorar as classes de produto mais compradas pelos lares, em bebida e comida, com o intuito de apurar o tipo de necessidades que precisam de ser satisfeitas.

Segue-se uma análise ao cabaz de compras, de acordo com a variedade de produtos comprados. Esta variável está intimamente relacionada com a dimensão das famílias existentes. Por exemplo, uma família composta por 5 membros compra, *a priori*, mais quantidade e, eventualmente, maior variedade de produtos, do que uma família com apenas 2 membros.

Para além destas duas análises, que irão complementar a análise exploratória, foi necessário identificar o tipo de frequência de compra em loja.

A amostra disponibilizada pela *Nielsen*, relativa aos 447 lares em análise, não pode ser expandida, de modo a representar o painel de lares, em Portugal Continental. Por isso, os dados em estudo podem não corresponder à realidade existente no país.

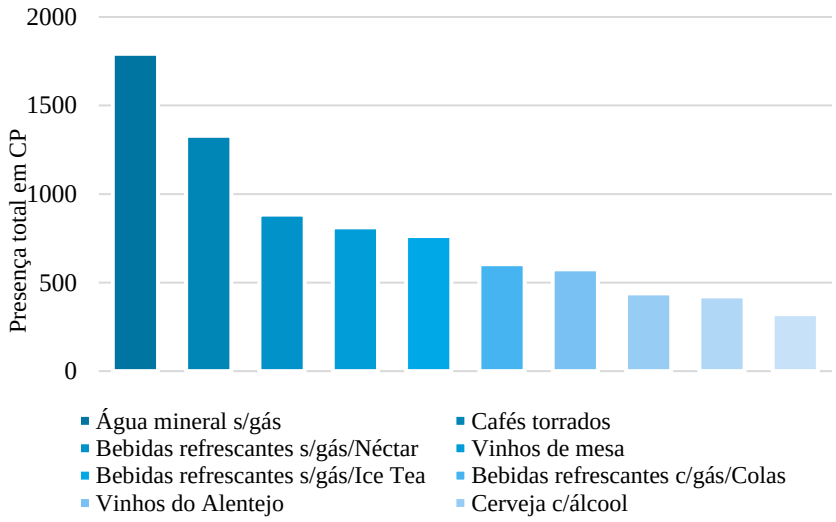

Gráfico 3.1.2.1 - Top de bebidas com maior presença nos cabazes dos lares

Os produtos de bebidas disponíveis para serem comprados pelos lares, encontram-se divididos por 44 classes de produto. No período compreendido entre 4 de janeiro e 17 de abril de 2016, estiveram presentes nos cabazes dos lares 10.775 bebidas. Estas bebidas, medidas em presenças, equivalem à presença em cabaz, ou seja, existe informação disponível de que a CP estava presente no cabaz de compras, não sendo possível apurar a quantidade e a diversidade de produto comprado. É possível concluir, através da leitura do Gráfico 3.1.2.1, que a Água mineral sem gás é a bebida com maior

representação nos cabazes dos portugueses, com 1.792 presenças. Seguem-se os Cafés torrados e as Bebidas refrescantes s/gás/Néctar, com 1.329 e 884 presenças, respetivamente.

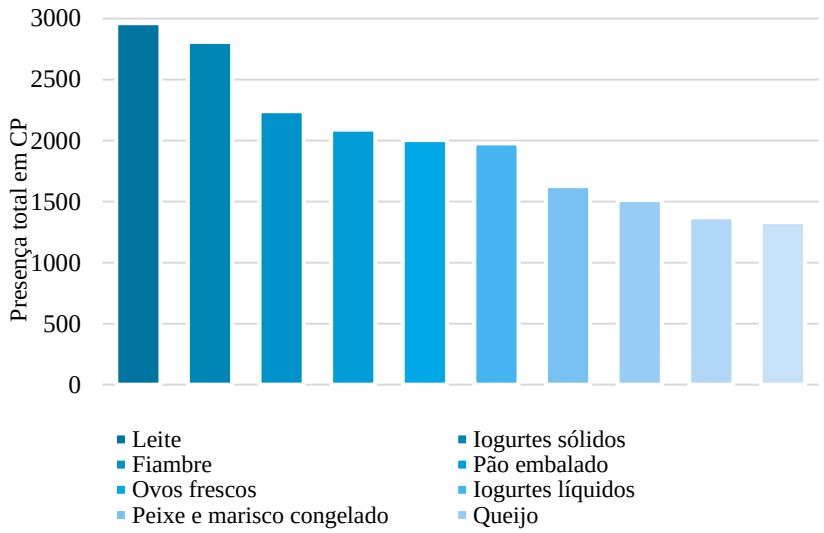

Gráfico 3.1.2.2 - Top de comidas com maior presença nos cabazes dos lares

Existem 137 classes de produto disponíveis, para dividir os produtos de comida, comprados pelos lares. Entre 4 de janeiro e 17 de abril de 2016, estiveram presentes, nos cabazes dos lares, 64.074 produtos de comida. A categoria dos produtos de comida é apresentada da mesma forma que a das bebidas, ou seja, é medida em presenças, o equivalente à presença em cabaz. Existe a informação disponível de que a CP estava presente no cabaz de compras, desconhecendo-se a quantidade e a diversidade de produto comprado.

É possível observar, no Gráfico 3.1.2.2, que o Leite é o produto de comida com maior presença no cabaz dos lares, em Portugal Continental, com 2.962 presenças. Seguem-se os Iogurtes sólidos e o Fiambre[15], com 2.808 e 2.241 presenças, respetivamente.

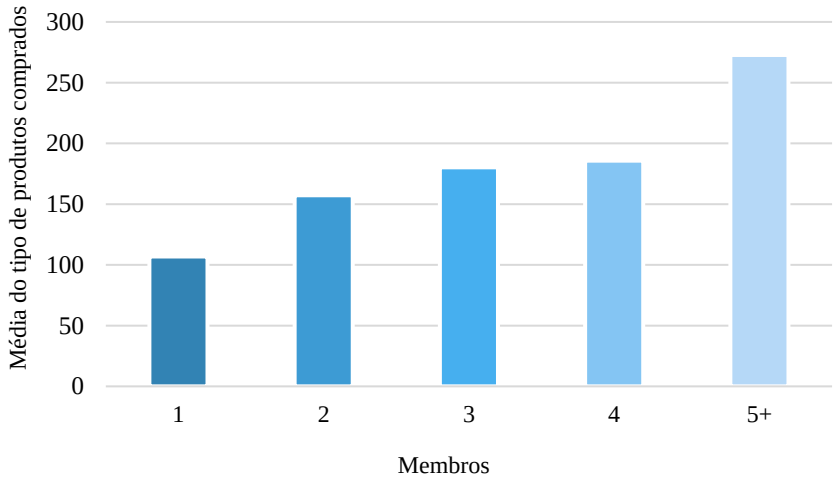

Gráfico 3.1.2.3 - Média da variedade de compras realizadas

[15] O produto Fiambre apenas é estudado no caso de embalagens de produto com peso fixo, não se considerando, desse modo, fiambre a peso, usualmente adquirido no serviço de charcutaria.

A partir do Gráfico 3.1.2.3, é possível analisar a média de presenças, por tipo de produto, dos lares. As famílias com apenas 1 membro compram, em média, 107 variedades de produto, ao longo do período de análise, enquanto as famílias com 2 e 3 membros, compram, respetivamente, 157 e 180 variedades de produto.

É, pois, possível concluir que a variedade do tipo de produtos comprados, em média, pelas famílias, está diretamente relacionada com o número de membros pertencentes a cada lar, isto é, quantos mais membros um lar tiver, maior será a variedade do tipo de produtos comprados.

A mesma análise foi realizada recorrendo à segmentação por tipo de família, não tendo sido possível observar uma ligação direta entre a variedade do tipo de produtos comprados e o tipo de família existente, uma vez que não foi possível diferenciar famílias Seniores de Famílias novas, pela variedade de classes de produto comprado, por exemplo.

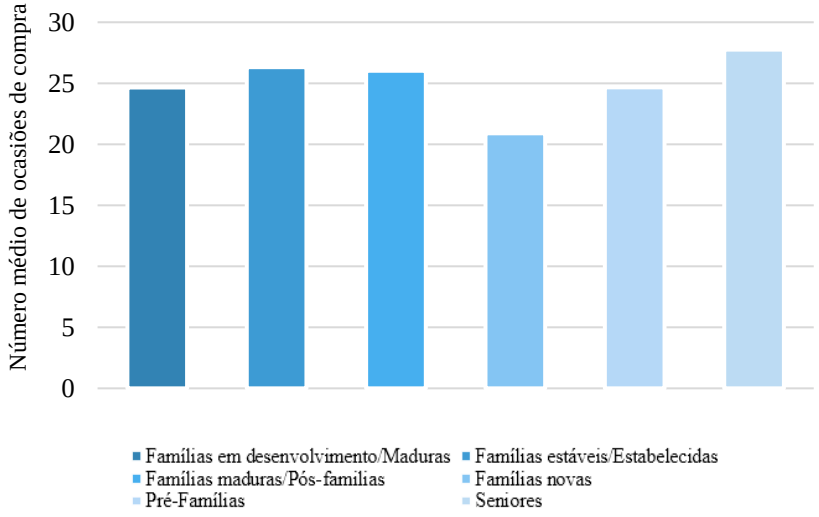

Gráfico 3.1.2.4 - Frequência média de compra realizada pelos lares

A frequência média de compra por lar foi de aproximadamente 27 ocasiões, ao longo dos 3,5 meses, para o total dos 447 lares. Ao longo de um mês, uma família foi aproximadamente oito vezes ao supermercado e o cabaz foi composto, em média, por sete variedades de produtos.

Através da representação do Gráfico 3.1.2.4, é possível observar que os Seniores foram o tipo de família que mais vezes foi às compras, no período em análise, com aproximadamente 27,8 idas, acompanhados de perto pelas Famílias estáveis/Estabelecidas e Famílias maduras/Pós-famílias, com 26,3 e 26 deslocações, respetivamente.

Por último, as Famílias em desenvolvimento/Maduras e as Pré-famílias deslocaram-se 24,7 vezes ao supermercado, enquanto as Famílias novas contabilizam um total de 20,9 deslocações.

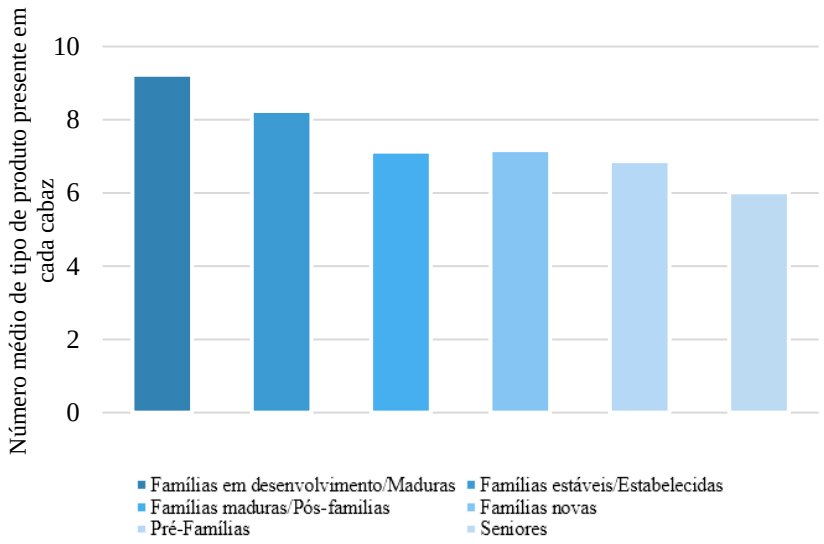

Gráfico 3.1.2.5 - Número médio de compras realizadas pelas famílias

Relativamente ao número médio de compras, por ocasião, por tipo de família, é possível observar, no Gráfico 3.1.2.5, que as Famílias em desenvolvimento/Maduras são aquelas que mais variedade de tipo produto compram, em cada ocasião, com 9,2 tipos de produto, apesar de serem o segundo tipo de família que menos vezes vai ao supermercado. Seguem-se as Famílias estáveis/Estabelecidas e as Famílias novas com 8,2 e 7,2, respetivamente. Os Seniores, apesar de serem os que contabilizam mais idas às compras, são quem compra menos quantidade de produto por ida, com um total de 6 tipos de produto, por cabaz, em média.

3.2. Pré-processamento dos dados

Para que o *raw data* pudesse ser analisado, através do programa *Rstudio*, foi necessário realizar um pré-processamento aos dados.

Os dados disponibilizados pela *Nielsen*, em bruto, foram recebidos através de um ficheiro *Excel*, em formato *.xlsx*.

	A	B	C	D	E	F	G	H	I	J	K
1	PERIOD	DATE	TIME	Lar	CP	PRODUCT	AREA	N. OF MEMBERS	CHILDREN	SOCIAL CLASS	TYPE OF FAMILY
2	2016001	20160104	1021	1	PC 0560 PEIXE E MARISCO CONGELADO	FOOD	AREA 1	2 MEMBROS	NO	MEDIA	SENIORES
3	2016001	20160104	1038	2	PC 0551 ALIMENTOS PARA CAES	FOOD	AREA 5	5+ MEMBROS	YES	MEDIA BAJA - BAJA	FAMILIAS EM DESENVOLVIMENTO - MADURAS
4	2016001	20160104	1059	3	PC 1600 OVOS FRESCOS	FOOD	AREA 5	2 MEMBROS	NO	MEDIA	SENIORES
5	2016001	20160104	1117	4	PC 0560 PEIXE E MARISCO CONGELADO	FOOD	AREA 2	3 MEMBROS	NO	MEDIA	SENIORES
6	2016001	20160104	1512	5	PC 0202 SOBREMESAS REFRIGERADAS	FOOD	AREA 1	3 MEMBROS	NO	MEDIA ALTA - ALTA	SENIORES
7	2016001	20160104	1700	6	PC 0551 ALIMENTOS PARA CAES	FOOD	AREA 4	4 MEMBROS	YES	MEDIA ALTA - ALTA	FAMILIAS ESTAVEIS - ESTABELECIDAS
8	2016001	20160104	1700	7	PC 0560 PEIXE E MARISCO CONGELADO	FOOD	AREA 3N	5+ MEMBROS	NO	MEDIA ALTA - ALTA	SENIORES
9	2016001	20160104	1701	8	PC 0660 REFEICOES PRONTAS CONGELADAS	FOOD	AREA 1	1 MEMBRO	NO	MEDIA	SENIORES
10	2016001	20160104	1729	9	PC 0070 CHAS	BEVERAGES	AREA 3S	4 MEMBROS	NO	MEDIA ALTA - ALTA	SENIORES
11	2016001	20160104	1800	10	PC 1600 OVOS FRESCOS	FOOD	AREA 3N	3 MEMBROS	NO	MEDIA	SENIORES
12	2016001	20160104	1808	11	PC 0710 COMPONENTES DE REFEICOES	FOOD	AREA 5	4 MEMBROS	YES	MEDIA	FAMILIAS EM DESENVOLVIMENTO - MADURAS
13	2016001	20160104	1822	12	PC 1600 OVOS FRESCOS	FOOD	AREA 2	2 MEMBROS	NO	MEDIA	FAMILIAS MADURAS - PÓS-FAMÍLIAS
14	2016001	20160104	1833	13	PC 0560 PEIXE E MARISCO CONGELADO	FOOD	AREA 1	2 MEMBROS	NO	MEDIA BAJA - BAJA	SENIORES

Figura 3.2.1 - Dados em bruto disponibilizados pela *Nielsen*

Foi necessário uniformizar, numa primeira fase, a coluna E, correspondente às classes de produto. Esta variável dispunha de demasiada informação que, para a análise desenvolvida, era dispensável. Assim,

as 181 classes foram reduzidas a 110, tendo sido depois agrupadas em classes de produto, que pertenciam à mesma família de produtos, através da criação de uma nova CP, mais abrangente.

	A
92	PC 0451 CONSERVAS DE PEIXE - ATUM POSTA
93	PC 0452 CONSERVAS DE PEIXE - ATUM SUB-PRODUTOS
94	PC 0453 CONSERVAS DE PEIXE - SARDINHAS
95	PC 0454 CONSERVAS DE PEIXE - CAVALA
96	PC 0455 CONSERVAS DE PEIXE - PASTAS
97	PC 0456 CONSERVAS DE PEIXE - LULAS
98	PC 0457 CONSERVAS DE PEIXE - OUTROS

	A
18	PC 0026 CONSERVAS DE PEIXE

Figura 3.2.2 - Exemplo de classes de produtos que foram agrupadas

Para esta análise, considerou-se irrelevante o tipo de conservas de peixe, por exemplo, que os consumidores compram, dando, sim, importância ao facto de os mesmos comprarem ou não conservas de peixe. O tipo de peixe adquirido pelo consumidor faz parte de uma outra análise, mais aprofundada, à preferência do consumidor, não abordada neste trabalho.

Nesse sentido, foi criado um novo ficheiro *Excel*, com os novos agrupamentos de CPs. Foram utilizadas as variáveis CP e lares, de modo a poder construir-se uma matriz. As classes de produto continham o nome por extenso, tendo-se optado por utilizar apenas a sigla CP, seguida de um número identificativo do tipo de classe – os valores foram escolhidos de forma aleatória, não correspondendo a nenhum número significativo.

B3 =LEFT(A3;FIND("";A3;7))

	A	B	C	D	E	F
1			Lares			
2	Lista de CPs		1	2	3	4
3	PC 0011 CAFES SOLUVEIS	PC 0011	1	0	0	0
4	PC 0012 CAPUCCINO (ESPECIALIDADES)	PC 0012	1	0	0	0
5	PC 0013 MISTURAS SOLUVEIS	PC 0013	0	0	0	0
6	PC 0014 SUCEDANEOS SOLUVEIS	PC 0014	0	0	0	0
7	PC 0015 SOBREMESAS	PC 0015	0	2	0	0
8	PC 0016 CHAS	PC 0016	3	0	3	2
9	PC 0017 BEBIDAS REFRESCANTES S/GAS	PC 0017	8	14	3	1
10	PC 0018 BOLACHAS	PC 0018	5	27	8	1
11	PC 0019 QUEIJOS	PC 0019	4	19	1	9

Figura 3.2.3 - Dados *raw data* em preparação

Na tabela apresentada na Figura 3.2.3, é possível observar como foram transformados os dados, através de uma *pivot table*, criada em *Excel*, tendo como objetivo tornar a leitura em *R* mais prática.

	A	B	C	D	E	F	G	H
1	Lar	PC 0011	PC 0012	PC 0013	PC 0014	PC 0015	PC 0016	PC 0017
2	1	1	1	0	0	0	3	8
3	2	0	0	0	0	2	0	14
4	3	0	0	0	0	0	3	3
5	4	0	0	0	0	0	2	1
6	5	0	0	0	0	2	1	0
7	6	0	0	0	1	1	1	1
8	7	1	3	2	1	0	4	0
9	8	1	3	0	0	1	1	5

Figura 3.2.4 - Matriz final

A Figura 3.2.4 disponibiliza os dados *raw data* das variáveis Lares e CP, convertidos numa matriz, de modo a poderem ser utilizados nas análises.

Após a execução destes passos, foi necessário gravar o ficheiro *Excel* em formato *.csv* (*Comma Delimited*), tendo a informação ficado pronta a inserir no *R*, através do seguinte comando:
read.csv ("Matriz_invertida_final.csv", sep =";", header=TRUE)

Porém, de modo a criar a matriz de dados em *R*, foi utilizado o seguinte comando:
matriz <- read.table("Matriz_invertida_final.csv", sep =";", header=TRUE)

matriz ×

Filter

	PC.0011	PC.0012	PC.0013	PC.0014	PC.0015	PC.0016	PC.0017
1	1	1	0	0	0	3	8
2	0	0	0	0	2	0	14
3	0	0	0	0	0	3	3
4	0	0	0	0	0	2	1
5	0	0	0	0	2	1	0
6	0	0	0	1	1	1	1
7	1	3	2	1	0	4	0
8	1	3	0	0	1	1	5
9	3	0	0	0	0	8	3

Figura 3.2.5 - *Data frame* da matriz

Desta forma, foi criado em *Rstudio*, um *data frame*, representado na Figura 3.2.5, construído a partir da combinação do número de lares e das CPs, tendo-se obtido uma matriz que demonstra a variedade, por tipo de produto, em ocasiões de compra, pelos diferentes lares.

A partir destes dados, foi ainda criada a matriz binária, através do seguinte comando:
BBin<-data.frame((matriz>0)*1)

BBin ×

Filter

	PC.0011	PC.0012	PC.0013	PC.0014	PC.0015	PC.0016	PC.0017
1	1	1	0	0	0	1	1
2	0	0	0	0	1	0	1
3	0	0	0	0	0	1	1
4	0	0	0	0	0	1	1
5	0	0	0	0	1	1	0
6	0	0	0	1	1	1	1
7	1	1	1	1	0	1	0
8	1	1	0	0	1	1	1
9	1	0	0	0	0	1	1

Figura 3.2.6 - *Data frame* da matriz binária

Esta matriz contém nas colunas as várias listas de compras e nas linhas a classe de produto, preenchida por 1, se o lar adquiriu o produto, ou 0, caso não o tenha feito.

Este pré-processamento foi realizado tendo como objetivo a identificação de perfis de consumo. Foram utilizadas a matriz binária, *BBin*, para a segmentação por presença de produto, no cabaz de compras, e a matriz invertida, matriz, para a análise por segmentação, por quantidade comprada.

No entanto, para a realização de *market basket analysis*, regras de associação entre os produtos alimentares, foi necessário adaptar a matriz binária, *BBin*, de forma a poder ser usada nesta análise. Optou-se por esta via, pois não é relevante a quantidade de produto comprada, mas sim o número de transações em que as CPs são compradas, de acordo com o número transações existentes. Se fosse considerada a quantidade de produtos, os padrões seriam afetados por um ato esporádico – *outliers*.

Para isso, procedeu-se, numa primeira fase, à instalação do *package arules*. De seguida, visto que a matriz binária foi criada em *dataframe*, foi necessário alterar o formato e torná-la matriz. Para esse efeito, foi utilizado o comando:
BBin<-as.matrix(BBin)

De acordo com o indicado no folheto do *package*, os dados devem ser introduzidos em formato de transações. Para converter os dados, utilizou-se o comando:
transactions<-as(BBin,"transactions")

Para verificar se os dados já se encontravam em condições de utilização, através da introdução de todas as CPs e transações, foi utilizado o comando:
inspect(transactions)

Assim, confirmou-se estarem reunidas as condições necessárias para a realização da análise de regras de associação, abordada mais à frente, no Capítulo 6.

Capítulo 4

Segmentação por presença de tipo de produto no cabaz de compras

Com o objetivo de se obter uma segmentação dos lares, tendo em conta a presença do tipo de produto comprado, procedeu-se a uma análise de *clusters*. O pré-processamento dos dados foi realizado no capítulo anterior e, por isso, estão reunidas as condições para iniciar a análise em *R*.

De seguida, é apresentado um subcapítulo, que contém uma breve discussão sobre o método de formação de *clusters* e o respetivo número associado, de acordo com o tipo de amostra existente. Depois, são apresentados os resultados obtidos, assim como a caracterização desses mesmos *clusters*, recorrendo à informação disponível sobre as variáveis. Por fim, procede-se à validação da análise de *clusters* efetuada.

4.1. Método de formação de *clusters*

O processo de criação de *clusters* divide-se em três passos essenciais. Primeiro, inicia-se o processo pela procura dos coeficientes de concordância ideais e pela forma de os calcular, com o objetivo de se construir a matriz de distâncias. De seguida, a criação da matriz de concordância $(D = 1 - D)$ e, por último, a escolha do agrupamento hierárquico. Esta metodologia é, no entanto, apenas válida para dados categóricos.

Inicia-se a procura por medidas de semelhança, para variáveis binárias, através do exemplo apresentado na Tabela 4.1.1, analisando o comportamento de três coeficientes[16] – coeficiente de concordância

[16] Note-se que a revisão teórica sobre este tema foi abordada no subcapítulo 2.3.2 - Medidas de proximidade.

simples (*matching coefficient*), *Jaccard* e *Sørensen-Dice*. O objetivo é encontrar e criar *clusters*, através do seu grau de semelhança, de acordo com as compras efetuadas em simultâneo, pelas famílias.

Tabela 4.1.1 - Exemplo de aplicação prática de três coeficientes

		Presenças de produto em diversas CPs					
Indivíduos	1	0	0	0	0	0	1
	2	1	1	0	0	0	0
	3	0	1	0	0	0	0

O coeficiente de concordância simples valoriza tanto a presença de compra comum entre dois indivíduos, como a ausência de compra do mesmo produto. Em baixo, são apresentados os resultados do grau de semelhança entre os indivíduos, considerando que os indivíduos $(1,2)$, de acordo com este coeficiente, são semelhantes em metade.

Através da observação da Tabela 4.1.1, é possível verificar que têm de semelhante a ausência de três produtos, não tendo, contrariamente a isso, nada de semelhante, pois não compram nenhum produto em comum. Os indivíduos $(1,3)$ não compram nada em comum, mas, de acordo com o coeficiente, são semelhantes em $^2/_3$. Por último, os indivíduos $(2,3)$ são iguais em $^5/_6$, tendo em comum um produto e quatro ausências. Assim, é possível notar que este coeficiente não oferece a solução pretendida de semelhança entre indivíduos, de acordo com as compras efetuadas.

$$s(1,2) = \frac{3}{6} = \frac{1}{2} \qquad s(1,3) = \frac{4}{6} = \frac{2}{3} \qquad s(2,3) = \frac{5}{6}$$

O seguinte coeficiente a ser analisado é o de *Jaccard*. Este valoriza a presença comum e as discordâncias entre dois indivíduos e descarta todas as ausências comuns. Assim, apresenta-se a distância entre os indivíduos, considerando que dois indivíduos são iguais, por ambos comprarem o mesmo produto.

$$s(1,2) = \frac{0}{3} = 0 \qquad s(1,3) = \frac{0}{2} = 0 \qquad s(2,3) = \frac{1}{2}$$

De acordo com este coeficiente, os indivíduos $(1,2)$ e $(1,3)$ têm grau de semelhança 0, ou seja, não têm nada de semelhante entre eles. Observando a Tabela 4.1.1, é possível notar que, de facto, os indivíduos não apresentam compras em comum. Em relação aos indivíduos $(2,3)$, estes têm um grau de semelhança de $^1/_2$, descartando todas as quantidades simultâneas de 0.

O último coeficiente, *Sørensen-Dice*, dá maior importância aos resultados de concordância de compra do produto, mas tem também em conta os resultados de discordância, $(1,0)$ e $(0,1)$, entre os indivíduos, para o cálculo da semelhança. Este coeficiente descarta os resultados de $(0,0)$, ausência simultânea, entre os indivíduos.

$$s(1,2) = \frac{0}{3} = 0 \qquad s(1,3) = \frac{0}{2} = 0 \qquad s(2,3) = \frac{2}{3}$$

Assim, para este coeficiente, os indivíduos $(1,2)$ e $(1,3)$ têm grau de semelhança 0. Efetivamente, os indivíduos não compraram nada em comum. Relativamente aos indivíduos $(2,3)$, estes têm $^2/_3$ de semelhança, ignorando as células em que o resultado obtido é $(0,0)$.

Este coeficiente, também, não aparenta ser o mais adequado, porque, apesar de não valorizar os resultados $(0,0)$ e de os descartar, dá um maior peso aos resultados $(1,1)$. Isto resulta numa desproporcionalidade entre os resultados de discordância e os de presença simultânea, embora seja importante ter em conta os resultados de $(1,0)$ e $(0,1)$.

Assim, foi depois necessário optar-se por um coeficiente que fosse ao encontro do objetivo pretendido: valorizar as compras das famílias portuguesas. O coeficiente de *Jaccard* aparentou ser o mais adequado, de todos os analisados. Segundo Everitt *et al.* (2011), para a aplicação da medida de similaridade *Jaccard*, pode utilizar-se a função *daisy*, no *package cluster*, do *software Rstudio*, através do seguinte comando:
d<-daisy (BBin, metric =c("gower"))

Esta função apresenta-se da seguinte forma:

$$s_{ij} = s(i,j) = \frac{\sum_{k=1}^{p} w(ij,k) \times s(ij,k)}{\sum_{k=1}^{p} w(ij,k)} \tag{4.1}$$

Em que $s(i,j)$ é uma média ponderada de $s(ij,k)$, com pesos $w(ij,k)$. Estes pesos tomam os valores 0 ou 1, dependendo se a comparação entre i e j for considerada válida. O valor de $w(ij,k)$ é definido como 0, se o resultado da variável k estiver ausente para um ou ambos os indivíduos i e j. Nas restantes situações, o valor é definido por 1.

Após a aplicação do coeficiente de *Jaccard*, no *Rstudio*, procede-se à mudança de variável na matriz de distâncias. Esta matriz apresenta-se como uma matriz de semelhanças, porém, para a análise em questão, o objetivo é encontrar a dissemelhança entre os lares e não o contrário.

Depois de eleito e implementado o coeficiente, escolhem-se os métodos a aplicar. Para a escolha do método de formação de *clusters*, de entre os dois métodos estudados, hierárquico e não hierárquico, deve optar-se pelo que nos dá a possibilidade de obter *clusters* naturais, vantagem que não é possível obter através do método não hierárquico, visto que requer, *a priori*, um número fixo de *clusters*. Desse modo, é escolhido o método hierárquico, que, por sua vez, se subdivide em dois tipos: aglomerativo e divisivo. De acordo com as características[17] sobre estes dois métodos, escolhe-se o aglomerativo – o mais comum neste tipo de análises –, pelo facto de o divisivo não conseguir recuperar tão facilmente de uma partição mal realizada e de ser, para além disso, menos eficiente, exigindo, à partida, uma maior capacidade computacional, visto que, na primeira divisão, necessita de ter em conta todas as possíveis divisões dos dados, em dois agrupamentos.

Após a escolha dos modelos, de acordo com o critério definido acima, decide-se o tipo de distância a utilizar, para a formação dos *clusters*. O método *single linkage* é um dos mais comuns, porém, um dos piores que pode ser utilizado neste caso, na medida em que o seu propósito é que a distância entre os vários *clusters* seja a menor possível. Contudo, o objetivo é criar grupos de *clusters* que sejam o mais diferentes possível, isto é, que as distâncias entre os vários grupos sejam as maiores possíveis. Deste modo, o método mais indicado, para este caso, é o *complete linkage*.

Assim, utilizando os métodos aglometativo e *complete linkage*, é criado o dendrograma, através do seguinte comando:

[17] Enunciadas na componente teórica do subcapítulo 2.3.3 - Métodos de formação de *clusters*.

```
T <- hclust(d, method ="complete", members=NULL)
plot (T, hang = 0.0)
```

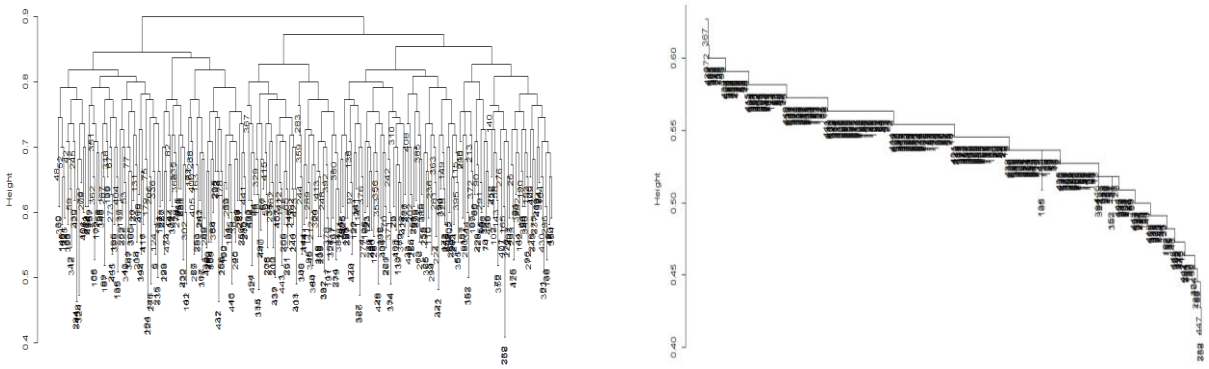

Figura 4.1.1 - Dendrogramas calculados com os métodos *complete* e *single linkage*

A partir dos dendrogramas apresentados na Figura 4.1.1, é possível concluir que o método *single linkage*, não é, tal como verificado anteriormente, o mais adequado, para a formação dos *clusters*. A distância entre dois grupos, utilizando o *single linkage*, é medida através do mínimo entre um par de indivíduos, entre todos os *clusters*. É, por isso, impossível, a partir deste método, conseguir dividir os dados em vários *clusters*. Por outro lado, no *complete linkage*, a medida apresenta-se como a distância máxima entre um par de indivíduos, entre todos os *clusters*. Assim, o objetivo passa por criar grupos de famílias que, entre si, tenham o maior número de hábitos de compra semelhantes e os mais diferentes possíveis, dos outros grupos.

Depois de criado o dendrograma, o passo seguinte recorre ao método que servirá de base, para obter o número de *clusters* ótimo. De acordo com Charrad *et al*. (2014), é possível utilizar o *package NbClust*, composto por trinta índices, para determinar o número de *clusters* ideal. Para isso, propõe ao utilizador o melhor esquema de agrupamento dos diferentes resultados obtidos, variando todas as combinações de número de *clusters*, medidas de distância e métodos de agrupamento.

O *package NbClust* é introduzido através do comando:
my.nbclust.object <-NbClust(data=NULL, diss=d, distance=NULL, method="complete", index=" ", min.nc=2, max.nc = 15)

Com isso, e de acordo com o *software Rstudio*, apenas os seguintes índices estão disponíveis*: Frey, McClain, Cindex, Silhouette* e *Dunn.*

Em baixo, são apresentadas todas as funções objetivo, aplicadas aos cinco índices disponíveis, de acordo com os parâmetros definidos anteriormente.

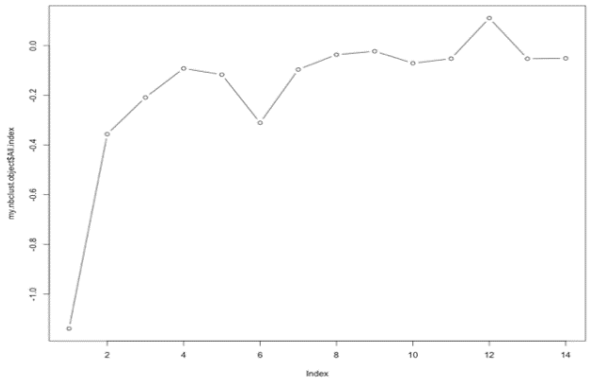

Figura 4.1.2 - Índice de *Frey* -
número de *clusters* vs. a função objetivo

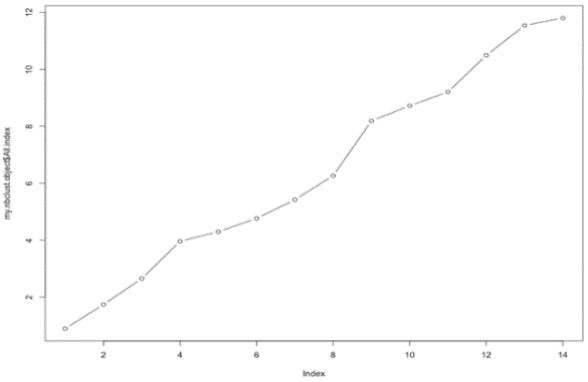

Figura 4.1.3 - Índice de *McClain* –
número de *clusters* vs. a função objetivo

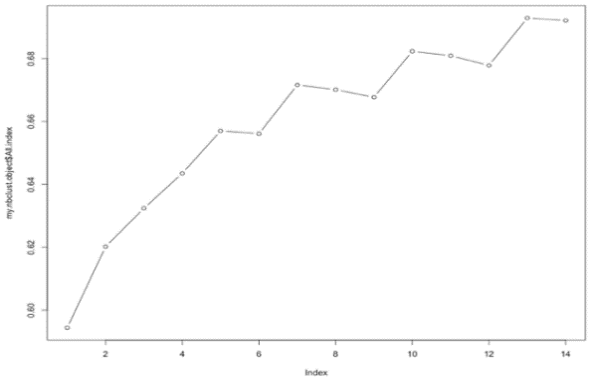

Figura 4.1.4 - Índice de *C-Index* -
número de *clusters* vs. a função objetivo

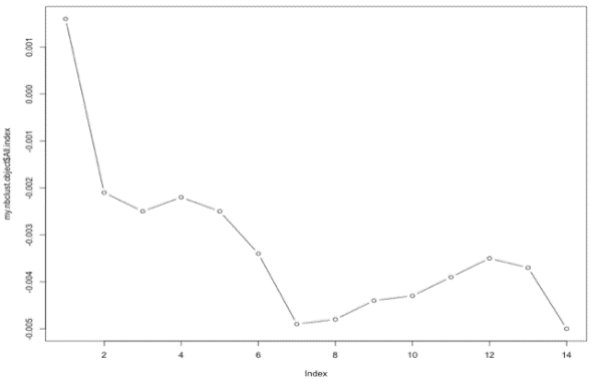

Figura 4.1.5 - Índice de *Silhouette* -
número de *clusters* vs. a função objetivo

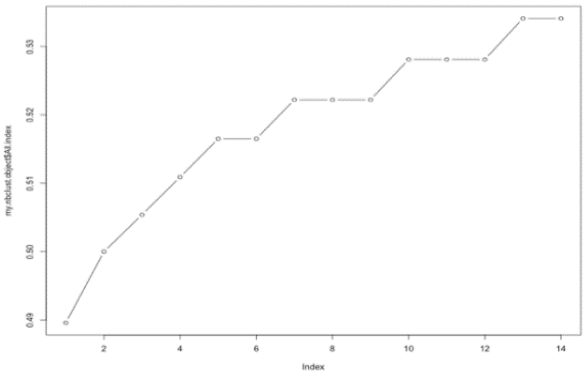

Figura 4.1.6 - Índice de *Dunn* -
número de *clusters* vs. a função objetivo

Na Figura 4.1.2, ilustrada pelo índice de *Frey*, aplicou-se o número de *clusters* pré-definido pelo *package*, de 2 a 15 *clusters*, assim como a todos os índices. Observando o comportamento da função, é possível verificar que não existe nenhum ponto que atinja a quantidade ótima definida pelos autores do modelo, 1,00, e, assim sendo, tal como indicado na teoria correspondente a este modelo, a solução de *cluster* único é assumida.

Segue-se o índice de *McClain*, apresentado na Figura 4.1.3. Aqui, é possível verificar que o rácio é sempre crescente, ou seja, o mínimo valor é atingido no primeiro cálculo, com dois *clusters*, obtendo-se um *index* de 0,8973 e, assim, a solução ótima do índice.

A Figura 4.1.4, corresponde ao índice *Cindex*, representado através de uma função crescente. Tal como no índice anterior, atinge o seu valor ótimo quando o mínimo valor da função é alcançado. Considera-se ótimo, quando a função atinge os dois *clusters*, com o valor de 0,5944.

O índice de *Silhouette*, ilustrado na Figura 4.1.5, é representado por uma curva decrescente, sendo que o seu valor ótimo é atingido quando o rácio é maximizado. Considera-se, assim, que o melhor valor atinge os dois *clusters*, com o rácio de 0,0016.

Por último, mas não menos importante, no índice de *Dunn*, apresentado na Figura 4.1.6, é possível observar que o gráfico mostra uma função crescente. Assim, e de acordo com o indicado pelo autor, o objetivo é maximizar o rácio, considerando-se 14 *clusters* como o valor ótimo, para se proceder ao corte dos dados, com o valor de índice de 0,5341.

Através da observação destes cinco índices, obtém-se o resumo do número ótimo de *clusters*, definido pelos modelos:

- 1 propôs 1 como o número ótimo de *clusters;*
- 3 propôs 2 como o número ótimo de *clusters;*
- 1 propôs 14 como o número ótimo de *clusters*.

De acordo com a regra da maioria absoluta, Charrad *et al*. (2014) definem, relativamente ao *package NcClust*, 2 como o número ótimo de *clusters*. Existe, ainda, outra opção que pode ser aplicada, considerando apenas os índices que preformam as melhores simulações em estudos. Milligan & Cooper (1985) estudaram as cinco melhores performances, correspondentes aos índices *CH index, Duda index, Cindex, Gamma* e *Beale*. Deste modo, é possível concluir, com bases nestes critérios, que o índice de *Cindex* é o modelo ótimo, para aplicar aos dados. Independentemente do critério utilizado, o número ótimo de *clusters* obtido é o mesmo – 2.

Analisou-se, ainda, o índice *PAM*, relativo ao modelo de partição, apresentado pelo *package pam*, com o seguinte comando:

```
P<-pam(d,2, diss=inherits(d,"dissimilarity"))
```

Este índice incorpora o modelo não hierárquico, tendo sido necessário fixar-se um número de *clusters* inicial. Nos modelos apresentados anteriormente, o valor ótimo foi 2 e, por isso, fixou-se igualmente 2, como o número de *clusters* indicado.

A Figura 4.1.7 representa o gráfico de *silhouette*, disponível no *package cluster*, onde é possível observar que o índice não apresenta bons resultados. Isto acontece, pois, este modelo não consegue

suportar *clusters* não esféricos e com tamanhos diferentes e conjuntos de dados dissimilares, como é o caso.

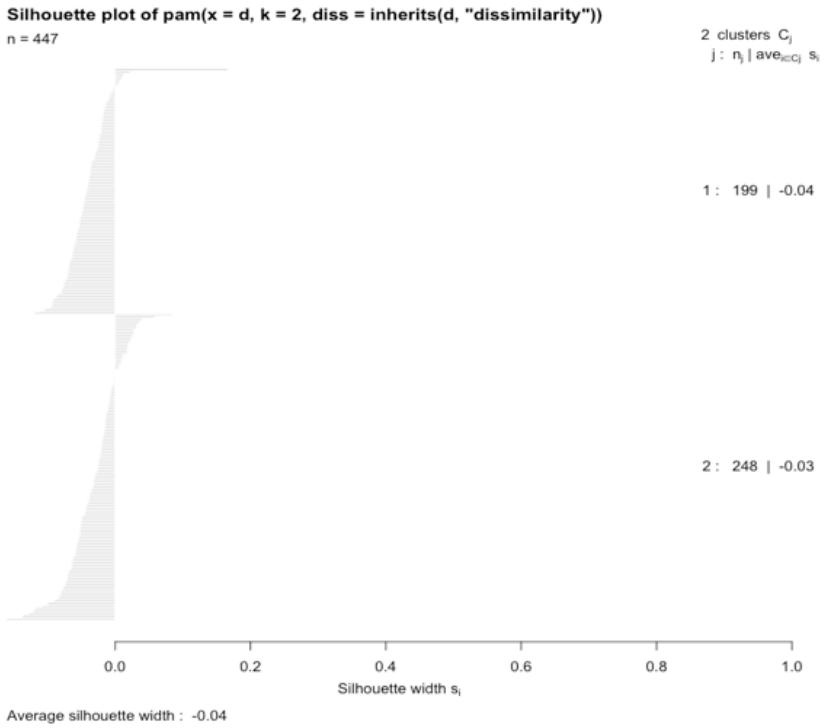

Figura 4.1.7 - Índice de *PAM* - número de *clusters vs.* a função objetivo

Desta forma, aplica-se o corte ao dendrograma, como é possível observar na Figura 4.1.8, realizado através do seguinte comando:
T.clust<-rect.hclust (T, k = 2, border = 2:5)

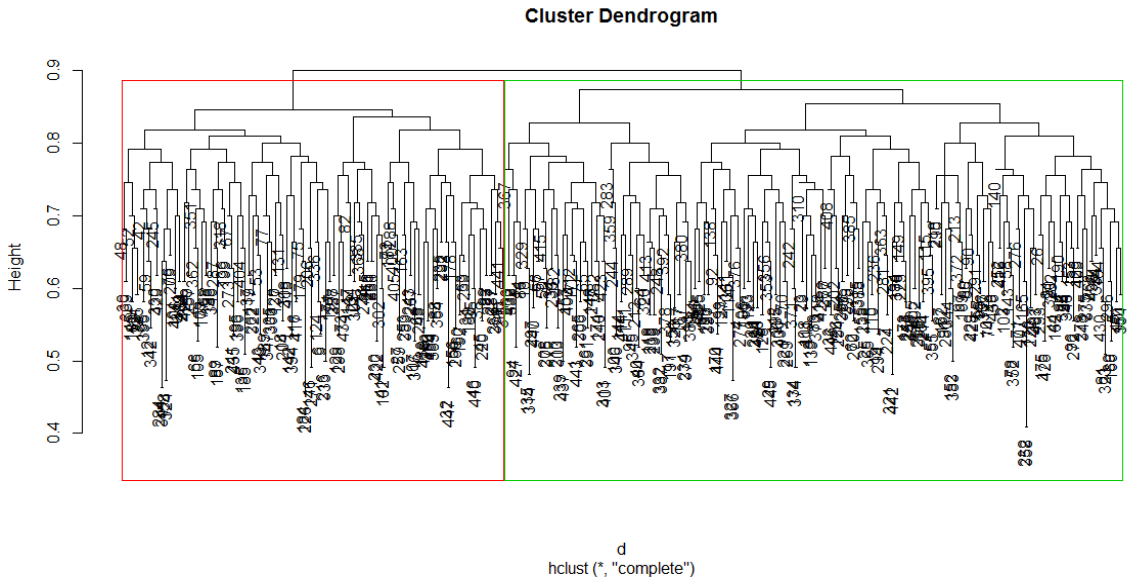

Figura 4.1.8 - Dendrograma dividido em dois *clusters*

De acordo com os índices estudados, conclui-se que o número de *clusters* ótimo é 2. Porém, observando o dendrograma representado na Figura 4.1.8, é possível verificar que existem mais possibilidades de corte, em 3 e 5 *clusters*. Assim, apesar de fixados os parâmetros necessários aos dados, a interpretação e caracterização, complementada com informação adicional sobre os lares, realiza-se para os três tipos de cortes mencionados, de forma a apurar se o corte ótimo, indicado pelos índices, é, realmente, o melhor.

4.2. Interpretação dos *clusters*

A interpretação dos *clusters* realiza-se em duas fases. A primeira inicia-se através da análise da Tabela 4.2.1, correspondente ao peso das características sociodemográficas, nos diferentes cortes aplicados aos lares, enquanto a segunda, relativa ao comportamento de compras, é efetuada a partir do estudo da presença de CP, no cabaz de compras.

Tabela 4.2.1 - Peso das características sociodemográficas nos *clusters*

		2 *Clusters*		3 *Clusters*			5 *Clusters*				
		C1	C2	C1	C2	C3	C1	C2	C3	C4	C5
Área de residência	1	0,34	0,28	0,34	0,24	0,30	0,36	0,32	0,24	0,26	0,35
	2	0,09	0,16	0,09	0,25	0,11	0,07	0,09	0,25	0,11	0,12
	3N	0,12	0,12	0,19	0,27	0,19	0,17	0,22	0,27	0,23	0,13
	3S	0,11	0,10	0,15	0,07	0,15	0,18	0,10	0,07	0,14	0,17
	4	0,19	0,21	0,12	0,08	0,14	0,12	0,13	0,08	0,19	0,07
	5	0,15	0,13	0,11	0,08	0,10	0,10	0,13	0,08	0,06	0,17
Números de membros por lar	1	0,10	0,15	0,10	0,17	0,15	0,12	0,08	0,17	0,18	0,11
	2	0,38	0,36	0,38	0,35	0,36	0,38	0,38	0,35	0,35	0,37
	3	0,33	0,27	0,33	0,30	0,26	0,27	0,42	0,30	0,22	0,31
	4	0,16	0,16	0,16	0,14	0,17	0,20	0,12	0,14	0,17	0,17
	5+	0,02	0,06	0,02	0,05	0,07	0,03	0.01	0,05	0,08	0,05
Tipo de família	Desenvolvimento/Maduras	0,11	0,11	0,11	0,13	0,10	0,11	0,12	0,13	0,07	0,13
	Estáveis/Estabelecidas	0,11	0,13	0,11	0,07	0,16	0,09	0,14	0,07	0,18	0,14
	Maduras/Pós-família	0,15	0,18	0,15	0,20	0,17	0,17	0,13	0,20	0,17	0,17
	Novas	0,04	0,05	0,04	0,05	0,05	0,04	0,04	0,05	0,05	0,05
	Pré-famílias	0,00	0,02	0,00	0,04	0,02	0,00	0,00	0,04	0,01	0,02
	Seniores	0,58	0,51	0,58	0,51	0,51	0,60	0,57	0,51	0,53	0,49
Presença de crianças	0	0,79	0,76	0,79	0,80	0,75	0,80	0,78	0,80	0,74	0,76
	1	0,21	0,24	0,21	0,20	0,25	0,20	0,22	0,20	0,26	0,24
Classe Social	Baixa/Média-Baixa	0,55	0,44	0,55	0,39	0,46	0,56	0,53	0,39	0,46	0,45
	Média	0,36	0,37	0,36	0,37	0,37	0,34	0,39	0,37	0,38	0,36
	Média-Alta/Alta	0,09	0,19	0,09	0,24	0,17	0,10	0,08	0,24	0,16	0,19

A interpretação dos dados é feita por característica sociodemográfica, como demonstra a Tabela 4.2.1, por forma a averiguar se existem diferenças de proporção, à medida que o número de *clusters* aumenta. Desse modo, e observando as proporções nas áreas de residência, para o *cluster* 2, é possível verificar que os dois grupos de lares são semelhantes. Relativamente aos grupos de 3 e 5 *clusters*, a semelhança não é evidente, ou seja, é possível observar as diferenças de proporções entre os grupos.

Para comprovar se as diferenças entre grupos são estatisticamente significativas, analisa-se a associação *interclusters*, através de um teste de qui-quadrado, testando a sua hipótese de independência. Se esta

hipótese for rejeitada, então os *clusters* são estatisticamente relacionados (Han *et al.*, 2011). Assim, é possível comprovar que a divisão em 3 e 5 *clusters* é benéfica e permite a criação de grupos distintos de lares, de acordo com as áreas de residência, com valores de *ρ-value* = 0,02 e 0,01, respetivamente.

Deste modo, o comportamento dos *clusters*, quando divididos em 3 grupos, transmite diferenças, que resultam em *clusters* heterogéneos, obtendo estes valores desiguais, principalmente nas áreas 1 e 2, correspondentes às áreas metropolitanas da Grande Lisboa e Porto, respetivamente. Ao mesmo tempo, há uma clara evidência de que as áreas 2 e 3N se encontram representadas pelo *cluster* 2, o mais dissemelhante entre todos. Por outro lado, é possível concluir, através da análise qui-quadrado, que os *clusters* 1 e 3 são os mais semelhantes.

A análise realizada a partir de 5 *clusters*, mostrou que, para além do *cluster* 3 ser o mais dissemelhante entre os grupos, apresenta, curiosamente, as mesmas proporções que C2, na análise a 3 *clusters*.

A caraterística Número de membros por lar, demonstra que famílias com 4 ou 5+ membros têm menor proporção, do que famílias com 2 ou 3 membros. Esta variável é semelhante para os três tipos de cortes, não existindo, assim, nenhuma evidência de mais-valia, em realizar o corte aos dados.

O tipo de família, no geral, não apresenta dissemelhanças *interclusters*, em todos os cortes, apesar de ser notório que vários *clusters* não são portadores de Pré-famílias. É, ainda, possível observar, na Tabela 4.2.1, que a maior fração provém de famílias Seniores, seguindo-se as Famílias maduras/Pós-famílias.

A presença de crianças é outra das características onde não é possível fazer-se a distinção entre *clusters*, sendo, as proporções, bastante semelhantes, em todos os grupos.

Por último, verifica-se, na Classe Social, dissemelhanças entre os *clusters* de 2 e 3. Isto significa que os rendimentos económicos de cada *cluster* são diferentes, existindo, por isso, *clusters* com maior capacidade de poder de compra, do que outros. O teste do qui-quadrado avalia esta associação, considerando-os diferentes entre si, com *ρ-value* = 0,01 em ambos os casos.

De acordo com o peso da característica Classe Social, com 2 *clusters*, é notória a diferença *interclusters*. No primeiro *cluster*, mais de metade dos lares representam a classe social Baixa/Média-Baixa, pertencendo, apenas, 0,1 à classe Média Alta/Alta. A classe social Média mantem-se semelhante para os dois *clusters*. O peso da classe social nos três *clusters* evidencia que a classe Média Alta/Alta é representada de forma igual. Relativamente às classes Baixa/Média-Baixa e Média, os *clusters* comportam-se de forma oposta, sendo de notar que a maior fração recai sobre a classe social Média, em todos os grupos em estudo.

Após uma análise às variáveis, com recurso ao teste qui-quadrado, conclui-se que a classe social Baixa/Média-Baixa é a mais dissemelhante, entre os três níveis apresentados.

Assim, e após serem apresentadas todas as características sociodemográficas sobre os lares, torna-se necessário investigar os comportamentos relativos aos hábitos de compra, através da análise dos cabazes, de cada *cluster*.

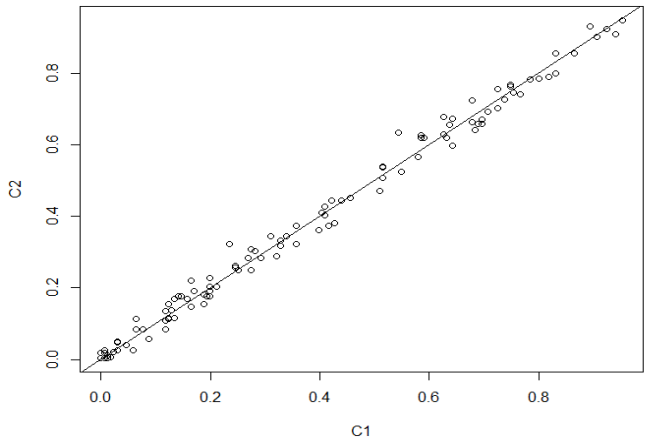

Figura 4.2.1 - Proporção dos *clusters*, C1 e C2, relativamente à presença de classe de produto no cabaz de compras

A Figura 4.2.1 demonstra a proporção de cada um dos *clusters*, relativamente à presença de classe de produto, no cabaz de compras. Desse modo, é possível verificar que ambos os *clusters* têm o mesmo comportamento, ou seja, a maioria das CPs encontra-se em cima ou bastante próxima da linha média. Assim, este tipo de corte do dendrograma não se torna uma mais-valia, quando o objetivo é analisar as presenças de CP no cabaz de compras de cada família. A mesma análise foi realizada nos casos em que o número de *clusters* é 3 ou 5. Porém, observa-se o mesmo comportamento, não existindo evidência de CPs que estejam mais presentes num *cluster* do que noutro, pelo que se conclui que o tipo de hábitos de consumo analisados não transmite diferenças significativas.

Assim, torna-se necessário realizar outra análise, tendo em conta a variedade de produtos comprados. Considera-se que a segmentação por presença de tipo de produto, no cabaz de compras, não é, por si só, uma análise suficientemente robusta, que permita a identificação de perfis de consumo, de bens de grande consumo, em lares portugueses, uma vez que não é possível obter semelhanças em praticamente nenhuma característica analisada, nos três tipos de cortes aos *clusters*.

Capítulo 5

Segmentação por quantidade comprada no cabaz de compras

Com o objetivo de segmentar os lares, tendo em conta a quantidade de produto comprado, realizou-se uma análise de *clusters*. O pré-processamento dos dados foi efetuado no capítulo 3, encontrando-se estes, por isso, em condições para iniciar o estudo.

Este capítulo contém uma breve discussão sobre o método de formação e o número de *clusters* ideal, de acordo com o tipo de amostra existente. De seguida, são apresentados os resultados obtidos, assim como a caracterização dos *clusters*, recorrendo à informação disponível sobre as variáveis.

5.1. Método de formação de *clusters*

O processo de criação de *clusters* divide-se em três passos essenciais. O primeiro pretende encontrar os coeficientes e calculá-los, de modo a construir a matriz de distâncias, através da medida de distância escolhida. De seguida, avança-se para a criação da matriz de concordância $(\mathrm{D} = 1 - \mathrm{D})$, e, por último, para a escolha do agrupamento hierárquico. Esta metodologia é, no entanto, apenas válida para dados categóricos.

Com a experiência obtida, através do capítulo anterior, relativamente à segmentação de *clusters*, por presença de tipo de produto, no cabaz de compras, prevê-se que algumas distâncias não sejam viáveis de aplicar aos dados. Algumas resultam no mesmo problema – a valorização da ausência de compra de produtos – e, desse modo, não fornecem os resultados pretendidos.

Desta forma, iniciou-se o processo de formação de *clusters*, através da escolha do coeficiente de concordância. Foram escolhidas e analisadas quatro distâncias, para serem aplicadas aos dados: distância *Euclidiana* e distância de *Gower* – ambas as distâncias previamente apresentadas no capítulo

anterior – e as distâncias *my.dist* e *my.dist.2* – abordadas de seguida. Para a escolha da distância[18], o objetivo resume-se a encontrar e criar *clusters* de famílias, através do seu grau de semelhança, de acordo com o que compram, em simultâneo.

Apresentam-se as distâncias, referidas acima, calculadas a partir da função *my.dist* e *my.dist.2*, que utilizam a distância *Euclidiana* com uma variante. Estas funções aplicam a distância *Euclidiana* dividida pelo número de observações em que não são ambas nulas.

Observa-se, através do seguinte comando, como são criadas as funções *my.dist* e *my.dist.2*, no *software Rstudio*:

```
my.dist<-function(x,y){
p<-length(x)
n.zeros<-sum(((x==0) & (y==0))*1)
sqrt(sum(x-y)^2)/(p-n.zeros)
}

my.dist.2<-function(x,y){
p<-length(x)
n.zeros<-sum(((x==0) & (y==0))*1)
sqrt(sum(x-y)^2/(p-n.zeros))
}
```

A matriz de distâncias é construída, de igual forma, para as duas medidas. Assim, *my.dist.D* recorre à função *my.dist* e *my.dist.D2* utiliza a função *my.dist.2* numa matriz de dados e constrói a matriz de distâncias.

De modo a compreender de que forma as distâncias, em cima referidas, operam, aplica-se um exemplo prático, através de uma matriz de dados. A matriz apresentada inclui quatro indivíduos, numerados de 1 a 4, e as suas quantidades compradas de quatro produtos, numerados de X_1 a X_4.

Desta forma, verifica-se que o produto 1 é comprado por todos os indivíduos, enquanto os restantes apenas são comprados pelos indivíduos 1 e 2.

	X_1	X_2	X_3	X_4
1	1	1	1	1
2	2	2	2	2
3	1	0	0	0
4	3	0	0	0

Assim, analisa-se, em detalhe, a equação utilizada para a criação da matriz de distâncias *my.dist.D*: $\frac{\sqrt{\sum (x-y)^2}}{(p-n.zeros)}$.

Na equação apresentada, designa-se por p o número de observações, sendo, para este exemplo, 4 e por $n.zeros$, o número de vezes em que ambos os indivíduos não compram o produto, ou seja, quando apresentam o valor 0. A distância de um indivíduo a outro ou vice-versa é a mesma, obtendo-se o mesmo valor para as duas situações. A distância entre o indivíduo a ele próprio é sempre 0.

[18] Note-se que a revisão teórica sobre este tema foi abordada no subcapítulo 2.3.2 - Medidas de proximidade.

A matriz de distâncias *my.dist.D*, apresenta-se calculada de seguida, através do *software* estatístico *Rstudio*. Desse modo, exibem-se os cálculos de forma a averiguar como são calculadas as distâncias entre indivíduos.

	[, 1]	[, 2]	[, 3]	[, 4]
[1,]	0.00	1.00	0.75	0.25
[2,]	1.00	0.00	1.75	1.25
[3,]	0.75	1.75	0.00	2.00
[4,]	0.25	1.25	2.00	0.00

A distância *my.dist.D* calcula a distância entre todos os indivíduos, da seguinte forma:

$$d_{1,2} = \frac{4}{4} = 1 \qquad d_{1,3} = \frac{3}{4} = 0{,}75 \qquad d_{1,4} = \frac{1}{4} = 0{,}25$$

$$d_{2,3} = \frac{7}{4} = 1{,}75 \qquad d_{2,4} = \frac{5}{4} = 1{,}25 \qquad d_{3,4} = \frac{2}{1} = 2$$

A distância entre indivíduos é calculada através da soma da diferença entre as unidades compradas por cada indivíduo, sendo, posteriormente, elevada ao quadrado e aplicada a raiz. De seguida, divide-se pelo número de produtos existentes, subtraindo-se pelo número total de zeros em comum entre os dois indivíduos, produtos não comprados, em simultâneo. Obtem-se, dessa forma, a associação entre dois indivíduos.

Observando a matriz de distâncias, é possível averiguar que não existem dois indivíduos com os mesmos comportamentos de compra. Isto indica que não existe nenhuma associação perfeita entre dois indivíduos.

Desta forma, por exemplo, a distância entre o indivíduo 1 e 2 apresenta o resultado 1, ou seja, os dois indivíduos diferem em 1 unidade de compra. Relativamente aos indivíduos 3 e 4, a sua distância é de 2, tendo em conta que os dois indivíduos apenas compraram um produto e que diferem de 2 unidades adquiridas.

Em relação à segunda matriz de distâncias, *my.dist.D2*, esta utiliza a distância *my.dist.2*, numa matriz de dados, e constrói a matriz de distâncias a partir da equação $\sqrt{\frac{\sum (x-y)^2}{(p-n.zeros)}}$, apresentada da seguinte forma:

	[, 1]	[, 2]	[, 3]	[, 4]
[1,]	0.00	2.00	1.50	0.50
[2,]	2.00	0.00	3.50	2.50
[3,]	1.50	3.50	0.00	2.00
[4,]	0.50	2.50	2.00	0.00

As distâncias são apresentadas da seguinte forma entre todos os indivíduos:

$$d_{1,2} = \frac{4}{2} = 2 \qquad d_{1,3} = \frac{3}{2} = 1{,}50 \qquad d_{1,4} = \frac{1}{2} = 0{,}50$$

$$d_{2,3} = \frac{7}{2} = 3{,}50 \qquad d_{2,4} = \frac{5}{2} = 2{,}50 \qquad d_{3,4} = \frac{2}{1} = 2$$

A distância entre indivíduos é calculada através da soma da diferença entre as unidades compradas por cada indivíduo, sendo, posteriormente, elevada ao quadrado. De seguida, divide-sem pelo número de produtos existentes, subtraindo-se o número total de zeros em comum, os produtos não comprados em simultâneo. No final, é aplicada a raiz quadrada a toda a função, obtendo-se, dessa forma, a associação entre dois indivíduos.

Observando a matriz de distâncias, é possível averiguar que não existem indivíduos com os mesmos comportamentos de compra, tal como na distância anterior, não existindo, assim, associações perfeitas entre dois indivíduos. A diferença entre as duas distâncias parte do local onde a raiz quadrada se encontra, incidindo no numerador, na primeira distância, e em toda a equação, na segunda. Os resultados obtidos na matriz de distância *my.dist.D2* são, maioritariamente, o dobro dos obtidos na matriz de distância *my.dist.D*, excetuando quando se calcula a distância entre os indivíduos 3 e 4.

Assim, por exemplo, as distâncias calculadas para os indivíduos 1 e 2 e 3 e 4 são iguais. Observando a matriz de dados, é possível verificar que os indivíduos 1 e 2 compram todos os produtos disponíveis, variando apenas nas unidades de compra, apresentando o resultado de 2. Os indivíduos 3 e 4 compram apenas um dos produtos disponíveis, mas em simultâneo, diferindo em 2 unidades, obtendo um resultado de 2.

A matriz de frequências é construída, de forma igual, para ambas as matrizes de distâncias, *my.dist.D* e *my.dist.D2*, e é apresentada através do seguinte comando, por exemplo, para a função *my.dist.D*:
d<-my.dist.D (matriz)
d<-as.dist(d)

Relativamente ao método de formação de *clusters*, considerou-se o método hierárquico, pois, com dados reais, não faria sentido definir-se, à partida, um número de *clusters*. Assim, optou-se pela sua formação natural, tal como no capítulo anterior. Entre os dois tipos de métodos hierárquicos, escolheu-se o método aglomerativo, através do critério *complete linkage*.

De seguida, são apresentados os quatro dendrogramas, construídos através das quatro distâncias referidas acima:

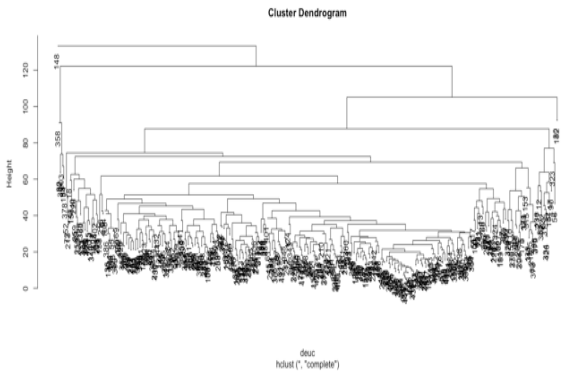

Figura 5.1 - Dendrograma utilizando a distância *Euclidiana*

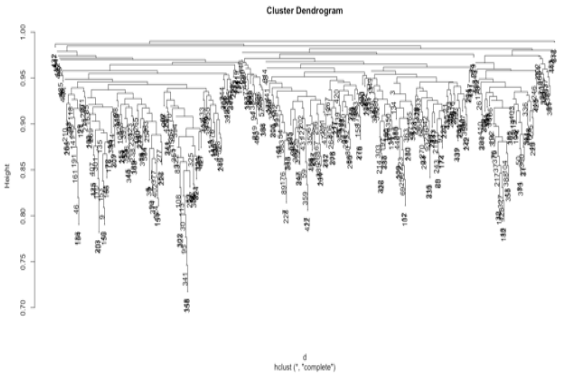

Figura 5.2 - Dendrograma utilizando a distância *Gower*

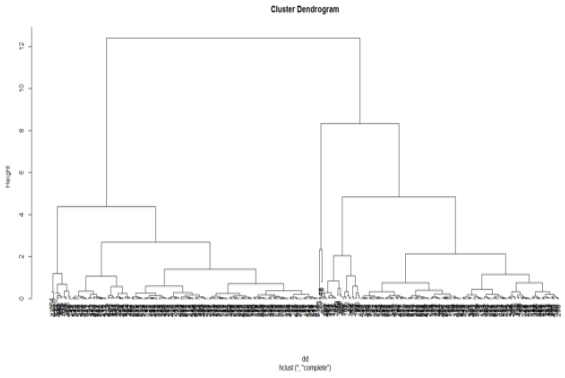

Figura 5.3 - Dendrograma utilizando a distância *my.dist*

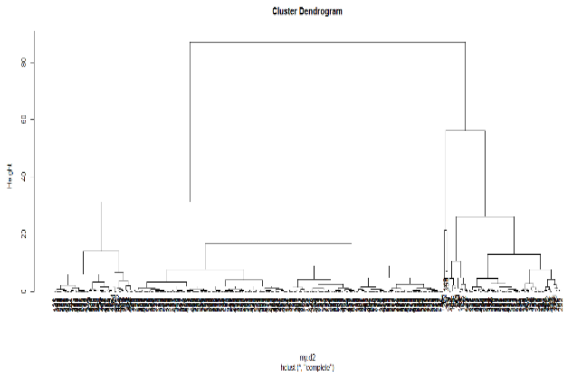

Figura 5.4 - Dendrograma utilizando a distância *my.dist2*

Os quatro dendrogramas apresentados acima, utilizam o método *complete linkage* na sua elaboração. Dessa forma, é possível observar que o Gráfico 5.1 e Gráfico 5.2, em que as distâncias *Euclidiana* e *Gower* são utilizadas, respetivamente, não representam bons resultados. Obtém-se uma composição descompensada e desproporcional, não sendo, por isso, viável realizar um corte no dendrograma, por forma a criar grupos de *clusters* homogéneos entre si, para a tomada de decisões.

As medidas formuladas especificamente para esta amostra, *my.dist* e *my.dist2*, apresentam dendrogramas bastante mais apelativos, Gráfico 5.3 e Gráfico 5.4, sendo possível, sem grande dificuldade, segmentar *clusters* e obter conjuntos de dados heterogéneos.

Deste modo, utilizam-se as duas medidas criadas e prosegue-se a análise, aplicando os cinco índices que estão disponíveis para serem aplicados aos dados, de acordo com os parâmetros definidos anteriormente, através do *package NbClust*: *Frey, McClain, Cindex, Silhouette* e *Dunn.*

A Tabela 5.1 apresenta os resultados obtidos para todos os índices, através das duas distâncias utilizadas.

Tabela 5.1- Resultados obtidos após aplicação do *package NbClust*

	my.dist		*my.dist2*	
	Número de *clusters*	Valor do índice	Número de *clusters*	Valor do índice
Frey	1	NA	1	NA
McClain	3	0,4046	3	0,2029
Cindex	2	0,1132	2	0,1334
Silhouette	3	0,5349	2	0,5792
Dunn	15	0,0353	15	0,0187

Tal como no capítulo anterior, aplicou-se o número de *clusters* pré-definido pelo *package*, de 2 a 15 *clusters*, a todos os índices. Através do índice de *Frey*, é possível afirmar que não existe nenhum ponto que atinja a quantidade ótima, definida pelos autores do modelo, seja 1,00 e, sendo assim, tal como indicado na teoria do modelo, a solução de *cluster* único é assumida. Esta situação acontece para as duas distâncias.

Segue-se o índice de *McClain*. A função objetivo para ambas as distâncias é semelhante pois o mínimo valor é atingido no segundo cálculo, ou seja, com três *clusters*, obtendo-se um *index* de 0,4046 e 0,2029, respetivamente –a solução ótima do índice.

O seguinte índice, correspondente ao *Cindex*, apresenta-se por uma função crescente, e tal como no índice anterior, atinge o seu valor ótimo quando o mínimo valor da função é alcançado. A função objetivo é idêntica nas duas distâncias, considerando-se o valor ótimo quando a função atinge os dois *clusters*, com os valores de 0,1132 e 0,1334, respetivamente.

O índice de *Silhouette*, representa-se por uma curva decrescente, sendo que o seu valor ótimo é atingido quando o rácio é maximizado. Este é o único índice que cria diferenças em relação ao número de *clusters* ideal, nas duas distâncias. Para a distância *my.dist*, considera-se que o melhor valor é alcançado com três *clusters*, obtendo-se o rácio de 0,5349. No caso da distância *my.dist.D*, o valor ótimo é atingido, quando a função objetivo obtém dois *clusters*, 0,5792.

Por último, o índice de *Dunn* é apresentado por uma função crescente. Assim, e de acordo com o indicado pelo autor do índice, o objetivo é maximizar o rácio. Consideram-se 15 *clusters* como o valor ótimo, para ambos os casos, de forma a proceder-se ao corte dos dados, com os valores de índice de 0,0353 e 0,0187, respetivamente.

Observando estes cinco índices, obtém-se o resumo do número ótimo de *clusters* definido pelos modelos para a distância *my.dist*:

- 1 propôs 1 como o número ótimo de *clusters;*
- 1 propôs 2 como o número ótimo de *clusters;*
- 2 propôs 3 como o número ótimo de *clusters;*
- 1 propôs 15 como o número ótimo de *clusters.*

Observando estes cinco modelos, obtém-se o resumo do número ótimo de *clusters* definido pelos modelos para a distância *my.dist2*:

- 1 propôs 1 como o número ótimo de *clusters;*
- 2 propôs 2 como o número ótimo de *clusters;*
- 1 propôs 3 como o número ótimo de *clusters;*
- 1 propôs 15 como o número ótimo de *clusters.*

Assim, de acordo com a regra da maioria absoluta, definida por Charrad *et al*. (2014), relativamente ao *package NcClust*, define-se 3 como o número ótimo de *clusters* para a distância *my.dist* e 2 o número ótimo para a distância *my.dist2*.

Apresenta-se a divisão das duas distâncias em dendrogramas, apresentando o corte ótimo definido pelo *package NbClust*.

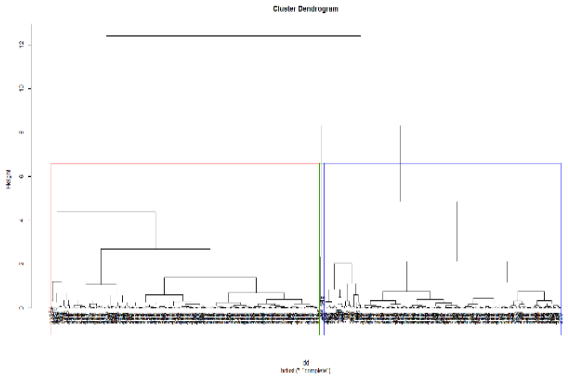

Figura 5.5 - Dendrograma apresentado o corte ótimo utilizando a distância *my.dist*

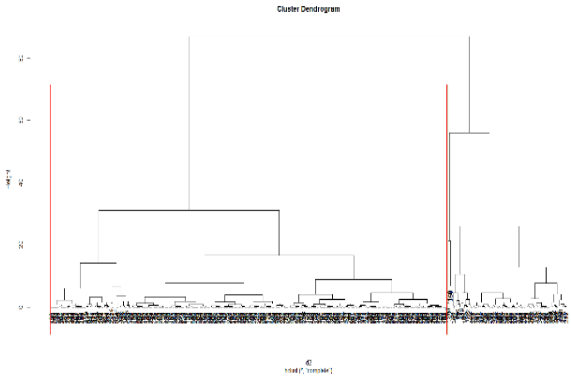

Figura 5.6 - Dendrograma apresentado o corte ótimo utilizando a distância *my.dist.2*

Analisa-se, ainda, o índice *PAM*, relativo ao modelo de partição, utilizado da mesma forma que no capítulo anterior. Este modelo incorpora o modelo não hierárquico, sendo necessário fixar um número de *clusters* inicial. Assim, define-se 3, como o número ótimo de *clusters* para a distância *my.dist*, e 2, como o número ótimo para a distância *my.dist.2*. Estes valores foram escolhidos com base no número de *clusters* ótimo, definido pelo *package NbClust*, de modo a fazer-se uma comparação entre índices.

Após a aplicação do índice de partição, obtém-se a Figura 5.7 e Figura 5.8, que apresentam os gráficos de *silhouette* dos dados, com a aplicação do modelo. É possível observar que os valores apresentam bons resultados, quando comparados com os resultados do capítulo anterior.

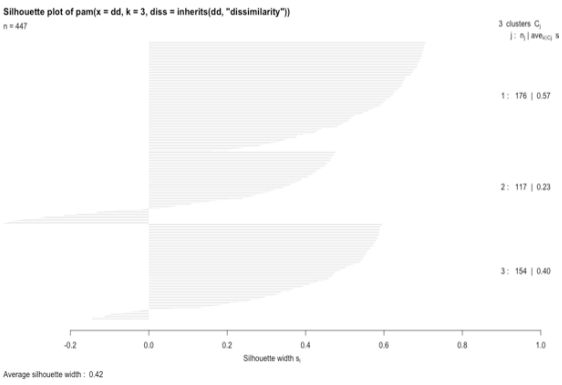

Figura 5.7 - Índice de *PAM* - número de *clusters* *vs.* a função objetivo aplicado à distância *my.dist*

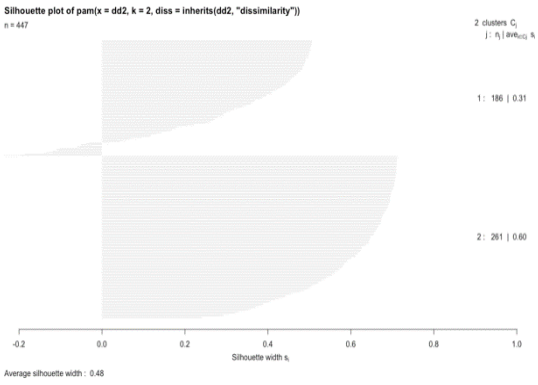

Figura 5.8 - Índice de *PAM* - número de *clusters* *vs.* a função objetivo aplicado à distância *my.dist2*

Deste modo, é necessário explorar os cortes aplicados aos dados, pelos dois índices, com as duas distâncias criadas, e compreender qual será o melhor índice e distância a utilizar, por forma a obter *clusters* heterogéneos. Assim, após fixados os parâmetros necessários aos dados, a interpretação e caracterização será complementada com informação adicional sobre os lares, por forma a perceber qual a distância e modelo que apresentam melhores *clusters*.

5.2. Interpretação dos *clusters*

A interpretação dos *clusters*, inicia-se através da Tabela 5.2.1. Deste modo, é possível observar como as famílias portuguesas são distribuídas pelos *clusters*, de acordo com a medida e o índice aplicado.

De acordo com o número de famílias presente em cada *cluster*, visualiza-se alguma desproporcionalidade nas suas dimensões. Os resultados obtidos pelo *package NbClust* refletem maior diferença, através do corte em 3 *clusters*, obtendo-se dois grupos com 235 e 208 famílias e um terceiro com apenas 4.

Tabela 5.2.1 - Resumo do número de famílias presente em cada *cluster*

NbClust					*PAM*				
2 *Clusters*		3 *Clusters*			2 *Clusters*		3 *Clusters*		
1	2	1	2	3	1	2	1	2	3
343	104	235	4	208	191	256	176	117	154

Desta forma, é necessário analisar em detalhe os *clusters* formados. Inicia-se a interpretação, através da análise da Tabela 5.2.2, que corresponde ao peso das características sociodemográficas nos diferentes *clusters*, através dos cortes aplicados aos lares. De seguida, observa-se o comportamento de compras dos *clusters*, através da proporção dos *clusters*, relativamente à presença de classe de produto no cabaz de compras, mas também a partir da frequência e variedade do cabaz de compras. Assim, será possível compreender as diferenças entre os grupos formados e qual a melhor distância e método a aplicar.

Tabela 5.2.2 - O peso das características sociodemográficas nos *clusters*

		NbClust					*PAM*				
		2 *Clusters*		3 *Clusters*			2 *Clusters*		3 *Clusters*		
		1	2	1	2	3	1	2	1	2	3
Área de residência	1	0,31	0,29	0,29	0,50	0,32	0,31	0,31	0,30	0,31	0,32
	2	0,13	0,13	0,14	0,00	0,12	0,12	0,12	0,12	0,13	0,11
	3N	0,21	0,18	0,20	0,50	0,20	0,20	0,19	0,22	0,18	0,19
	3S	0,12	0,18	0,12	0,00	0,16	0,15	0,13	0,13	0,11	0,17
	4	0,13	0,12	0,15	0,00	0,10	0,10	0,15	0,12	0,15	0,10
	5	0,11	0,11	0,11	0,00	0,10	0,11	0,10	0,11	0,11	0,10
Números de membros por lar	1	0,17	0,01	0,23	0,00	0,03	0,02	0,19	0,09	0,24	0,01
	2	0,38	0,32	0,38	0,00	0,35	0,32	0,39	0,38	0,39	0,29
	3	0,29	0,33	0,25	0,25	0,35	0,34	0,28	0,32	0,24	0,34
	4	0,14	0,22	0,13	0,00	0,20	0,20	0,14	0,17	0,12	0,21
	5+	0,02	0,13	0,01	0,75	0,07	0,12	0,01	0,03	0,01	0,14
Tipo de família	Desenvolvimento/Maduras	0,08	0,20	0,06	0,25	0,16	0,18	0,07	0,12	0,05	0,19
	Estáveis/Estabelecidas	0,11	0,16	0,09	0,25	0,16	0,17	0,11	0,15	0,06	0,18
	Maduras/Pós-família	0,16	0,18	0,13	0,25	0,21	0,18	0,15	0,16	0,18	0,18
	Novas	0,05	0,03	0,06	0,00	0,03	0,02	0,06	0,05	0,04	0,03
	Pré-familias	0,02	0,00	0,02	0,00	0,00	0,00	0,02	0,02	0,01	0.01
	Seniores	0,57	0,42	0,63	0,25	0,44	0,44	0,59	0,51	0,66	0,41
Presença de crianças	0	0,81	0,67	0,82	0,75	0,72	0,71	0,81	0,75	0,87	0,69
	1	0,20	0,33	0,18	0,25	0,28	0,29	0,19	0,25	0,13	0,31
Classe social	Baixa/Média-Baixa	0,16	0,14	0,17	0,25	0,13	0,14	0,16	0,15	0,17	0,14
	Média	0,46	0,54	0,47	0,25	0,50	0,49	0,48	0,44	0,50	0,52
	Média-Alta/Alta	0,38	0,32	0,36	0,50	0,37	0,36	0,37	0,41	0,32	0,34

A interpretação dos dados é feita por característica sociodemográfica, como demonstra a Tabela 5.2.2, de forma averiguar se existem diferenças de proporção, de acordo com as duas distâncias e índices utilizados. Inicia-se a análise, comparando a mesma distância, aplicada a modelos diferentes, investigando qual a que representa, na sua maioria, *clusters* heterogéneos.

Desse modo, observando as proporções na Área de residência, para a distância *my.dist*, é possível verificar que o modelo *PAM* cria grupos com características semelhantes, não existindo praticamente diferenças. Relativamente aos resultados obtidos a partir do *package NbClust*, observam-se valores bastante diferentes. É possível verificar que o grupo que contém apenas quatro famílias, quando comparado aos restantes *clusters*, apresenta diferenças bastante significativas, criando muitas variáveis com valores 0.

Para a segunda distância apresentada, *my.dist2*, não se observam diferenças significativas, considerando que os *clusters* são homogéneos. Assim, não há evidências de mais-valia pela utilização de qualquer um dos modelos.

Em relação à característica Número de membros por lar, a divisão relativa a 3 *clusters*, com recurso ao *package Nbclust*, é heterogénea. Praticamente não existem semelhanças *intracluster*, obtendo-se valores bastante diferentes em quase todas as dimensões de famílias, exceto quando os lares são constituídos por 3 membros.

Com a utilização do modelo *PAM*, em relação à formação em 3 *clusters*, observam-se, em geral, bastantes semelhanças. Apenas duas características se destacam, quando a característica Número de membros por lar tem a dimensão 1 e 5+, os extremos deste indicador.

A formação, em dois *clusters*, obtida através dos dois modelos aplicados, não apresenta diferenças significativas, tendo exatamente as mesmas dissemelhanças que o caso analisado em cima. Verificam-se, deste modo, disparidades relativas à dimensão por 1 e 5+ membros.

O tipo de família, no geral, não revela grandes dissemelhanças *interclusters*, apesar de ser notório que vários *clusters* não são portadores de Pré-famílias, por exemplo. Analisando a distância *my.dist*, corte em 2 *clusters*, e independentemente do modelo aplicado, os resultados obtidos são semelhantes, encontrando-se diferenças nas Famílias em Desenvolvimento/Maduras e Pré-famílias.

Em relação à distância *my.dist.2*, corte em 3 *clusters*, as diferenças entre a utilização dos índices varia. Apresentam-se, com a utilização do *package NbClust*, diferenças em todos os parâmetros, sobresaindo, tal como nas outras características, o C2 como o mais dispare entre todos. A utilização do índice *PAM*, apresenta, também, algumas diferenças, sendo, porém, menos notórias, do que no anterior índice aplicado, considerando-se, igualmente, C2 como o *cluster* mais dissemelhante.

Através da análise a característica Presença de crianças, não é possível fazer a distinção entre *clusters*, uma vez que as proporções são bastante semelhantes, em todos os grupos de *clusters*, e utilizam todas as distâncias e modelos.

Observando, de forma geral, como se comporta a característica Classe Social, é possível verificar que não existem praticamente dissemelhanças entre os *clusters* formados, através das duas distâncias criadas e dos modelos utilizados, com expeção do grupo de 4 famílias criado.

Assim, conclui-se que, de facto, existem diferenças pontuais ao analisar-se a formação dos *clusters*, com recurso às características sociodemográficas. Porém, através das duas medidas criadas e dos modelos aplicados, ainda não é possível concluir qual o melhor cenário a aplicar.

Deste modo, e após serem apresentadas todas as características sociodemográficas sobre os lares, torna-se necessário investigar os comportamentos relativos aos hábitos de compra, através da análise aos cabazes, de cada *cluster*, e à frequência média.

Tabela 5.2.3 - Frequência média, em dias, e variedade de cabaz de compras aplicados às duas distâncias e modelos em estudo

	NbClust					*PAM*				
	2 *Clusters*		3 *Clusters*			2 *Clusters*		3 *Clusters*		
	C1	C2	C1	C2	C3	C1	C2	C1	C2	C3
Frequência média de compra	23,7	6,0	22,0	60,8	31,2	32,2	9,0	26,8	19,6	35,5
Quantidade de produtos comprados por cabaz de compras	36,1	9,8	5,2	22,4	8,7	22,4	5,3	6,9	4,8	9,7

A Tabela 5.2.3, apresenta os resultados relativos à frequência média e quantidade de produtos comprados por cabaz, para os vários *clusters*. Se, em termos de características sociodemográficas, não se verificavam grandes dimissilhanças entre os *clusters*, nesta análise, é possível averiguar que, estes dois tipos de comportamentos, mostram *clusters* heterogéneos, com diferenças bastante significativas.

Analisando, de forma detalhada, a distância *my.dist*, repartida em 2 *clusters*, através do *package NbClust*, é possível concluir que existem grandes diferenças *interclusters*. As famílias representadas no C1, vão, em média, 24 vezes ao supermercado nos 105 dias de observação, enquanto as famílias do C2 têm uma frequência média de 6 vezes. Em relação à quantidade de CPs compradas por ato de compra, observa-se, uma vez mais, uma grande disparidade. Os lares de C1 compram, em média, 36 produtos por cada ida à compras, enquanto os lares de C2 compram aproximadamente 10 produtos distintos.

Realizada a mesma análise, utilizando o índice *PAM*, verificam-se, também, valores díspares, tanto na frequência de compra média, como em quantidade de CPs presentes no cabaz de compras. Em relação ao C1, este conjunto de famílias desloca-se, em média, 32 vezes ao supermercado, durante os 3,5 meses, enquanto no C2 isso acontece, apenas, 9 vezes. Na quantidade de produtos comprados, por cabaz de compras, C1 atinge uma quantidade média de 22 produtos e o C2 de aproximadamente 5 produtos.

Observando a formação dos dados, em 3 *clusters*, utilizando o *package NbClust*, verifica-se, uma grande dispersão nos valores apresentados. Considera-se que C2 engloba um grupo de famílias *outlier*, pela sua grande diferença de valores, em comparação com os restantes *clusters*. Este conjunto de famílias, C2, apresenta um elevado valor de frequência de compra, com 61 idas ao supermercado, em 105 dias de análise. Os restantes *clusters*, C1 e C3, vão respetivamente, 22 e 31 vezes. Relativamente à quantidade de produtos comprados, em cabaz de compras, o resultado do C2 é bastante mais elevado, comprando este, em média, 22 produtos diferentes em cada ida ao supermercado. Os restantes *clusters* compram 5 e 9 produtos, aproximadamente.

Utilizando o índice de *PAM* para o corte em 3 *clusters*, obtiveram-se, também, *clusters* heterogéneos. Porém, não tão evidente como no índice anterior.
Por norma, as famílias que vão mais regularmente às compras, adquirem, em condições normais, menos quantidade de produto do que famílias que vão apenas algumas vezes por mês. Neste caso, avaliando os

resultados obtidos, apresent-se o caso oposto. O C3 representa o conjunto de famílias que mais vezes se desloca às compras – cerca de 36 vezes no período em análise –, adquirindo, em média, 10 produtos. O C2 é o que vai menos vezes às compras – cerca de 20 vezes –, nos 105 dias em análise, tendo uma quantidade de produtos comprada de 5.

Em suma, é possível verificar que a criação de *clusters* naturais, a partir de segmentação por quantidade comprada no cabaz de compras, continua a não apresentar bons resultados, quando analisadas as características sociodemográficas dos 447 lares em estudo. Não é possível perceber qual o critério utilizado para a segmentação dos lares, a partir das duas distâncias criadas, *my.dist* e *my.dist2*, e dos dois índices utilizados, *NbClust* e *PAM*, pelo facto de não se identificarem *clusters* heterogéneos.

Desse modo, analisou-se o comportamento de compra das famílias, interpretando as várias segmentações existentes. Nesta segunda análise, já foi possível observar algumas dissemelhanças, destacando-se a divisão em 3 *clusters*, utilizando o *package NbClust*. Porém, de qualquer forma, com apenas estes indicadores, não é possível identificar o melhor cenário a aplicar aos dados.

Capítulo 6

Regras de associação aos cabazes de compras

De forma a definir os parâmetros para a realização da *market basket analysis*, torna-se, em primeiro lugar, necessário ter em conta que o estudo dos cabazes de compras irá incidir no total de transações realizadas, pelos lares, em 3,5 meses, não se podendo ter a perspetiva do cabaz de compras associado a cada ida ao supermercado. Isto significa que os valores, para *support* e *confidence*, serão mais elevados do que o habitual. Estando perante produtos alimentares, é importante ter em consideração que a rotatividade[19] de alguns deles é bastante elevada e que há artigos presentes em quase todos os cabazes de compras, porque se tratam de produtos que a maioria dos consumidores compra, como leite. Assim, este aspeto limita a análise, na medida em que não se pode ter a perspetiva das associações presentes em cada cabaz simples.

6.1. Definição do valor dos parâmetros

Analisa-se, de uma forma geral, como é que as várias transações – neste caso, as 447 famílias – se comportam, nas 110 CPs, ao longo do período em análise.

Através da Figura 6.1, é possível observar a frequência com que as diversas CPs são compradas, em relação às 447 transações. Desse modo, verifica-se que o cabaz de compras das famílias portuguesas inclui uma variedade de CPs. Porém, desconhece-se se as compras são esporádicas ou frequentes. Torna-se, por isso, pertinente analisar a frequência de compra dos produtos e as regras de associação existentes.

[19] Rotatividade de um produto consiste na medição da frequência com que um produto é comprado, num determinado período.

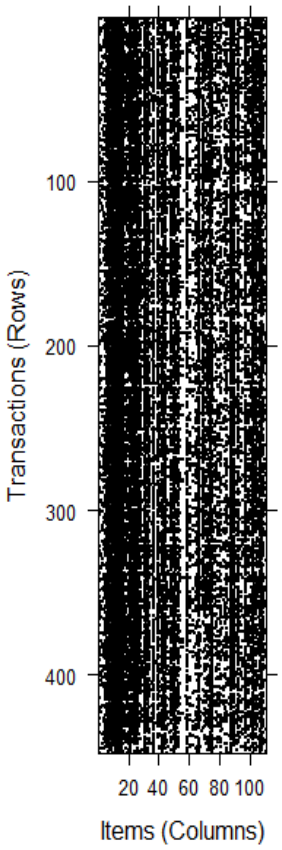

Figura 6.1 - Representação de todas as CPs compradas

Para definir os valores dos parâmetros, foi necessário considerar todos os níveis da hierarquia. É esperado que se obtenham valores mais elevados para os parâmetros dos níveis mais simples, do que para os mais complexos. Torna-se, por isso, essencial ter em atenção os valores escolhidos, a aplicar nos parâmetros, de forma a não excluir demasiados *itemsets*, mas também a não validar todas as regras.

Em relação ao nível de *confidence*, este fator terá um valor baixo, pois serão escolhidas as regras de associação de acordo com o *lift*. Este parâmetro, *lift*, representa a predisposição de um consumidor para adquirir um determinado artigo, aquando da compra de outro. Utiliza-se a validação com o parâmetro *lift*, com o objetivo de eliminar regras, nos artigos em que a sua compra simultânea seja aleatória, ou seja, por puro acaso.

Através do seguinte comando, identificam-se os *itemsets* frequentes e derivam-se as regras de associação, a partir do algoritmo *Apriori*, de acordo com os valores, para os parâmetros, *support* e *confidence*, definidos. Tendo sido utilizada a matriz binária *trans*.

```
basket_rules <- apriori(trans,parameter = list(sup = 0.65, conf = 0.75,target="rules"))
```

Deste modo, são criadas duas tabelas para testar a sensibilidade do parâmetro *support*. Na Tabela 6.1, são apresentados os *itemsets* frequentes, construídos a partir da fixação dos parâmetros 0,65, 0,75 e 0,85, para o parâmetro *support*. Na Tabela 6.2, impõe-se o valor de *confidence*, em 0,75, e varia-se o *support* entre os três valores apresentados, de onde se obtêm as regras de associação derivadas para 4 níveis hierárquicos.

Tabela 6.1 - Teste à sensibilidade do parâmetro *support*

Itemsets frequentes				
	Nível			
Support	1	2	3	4
0,65	28	220	522	328
0,75	7	58	78	16
0,85	7	16	0	0

Tabela 6.2 - Teste à sensibilidade do parâmetro *support* com nível de *confidence* superior a 0,75

Regras de associação				
	Nível			
Support	1	2	3	4
0,65	9	161	460	326
0,75	7	56	78	16
0,85	7	16	0	0

Na Tabela 6.1, é testada a sensibilidade do parâmetro *support* e é apresentado o número de artigos presentes em cada nível hierárquico do *itemset*. Assim, é possível observar que na base de dados disponibilizada, das 110 CPs existentes, apenas 28 têm *support* superior a 0,65, ou seja, estão presentes em 65% ou mais nos cabazes dos portugueses. Dos $\binom{28}{2}$ = 378 possíveis pares, constituídos para artigos de nível hierárquico 2, apenas 220 ocorrem com frequência de 0,65.

Como esperado, em qualquer uma das tabelas em análise, à medida que o valor de *support* aumenta, o número de *itemsets* frequentes e regras de associação diminuem, respetivamente. É, também, possível observar que o número de *itemsets* frequentes é maior do que o número de regras de associação. Isto acontece pelo facto de se ter estabelecido o valor de 0,75, para *confidence*.

Analisando os valores em que o parâmetro *support* é 0,85, nas duas tabelas, verifica-se a mesma quantidade. Esta situação ocorre, pois, esses mesmos *itemsets* e regras de associação, têm o valor de *support* superior a 0,85 e valor de *confidence* superior a 0,75.

De acordo com os resultados obtidos, e de forma a poderem obter-se regras de associação em todas as hierarquias, fixa-se os valores de *support* e *confidence*, em 0,75. Porém, interessa explorar um pouco os *itemsets* criados, para verificar se todos eles são pertinentes de análise.

6.2. *Itemsets* frequentes para cada nível hierárquico

Através da Figura 6.2, apresenta-se a frequência de compra, com parâmetro *support* de valor superior ou igual a 0,8, obtendo-se, assim, 9 artigos frequentes. Desta forma, interessa entender se, de facto, faz sentido a criação de regras de associação para artigos que estão, em alguns casos, presentes em mais de

95% dos cabazes de compras. Para isso, torna-se necessário compreender se as regras presentes são pertinentes para a análise, ou seja, se será possível encontrar associações interessantes entre artigos.

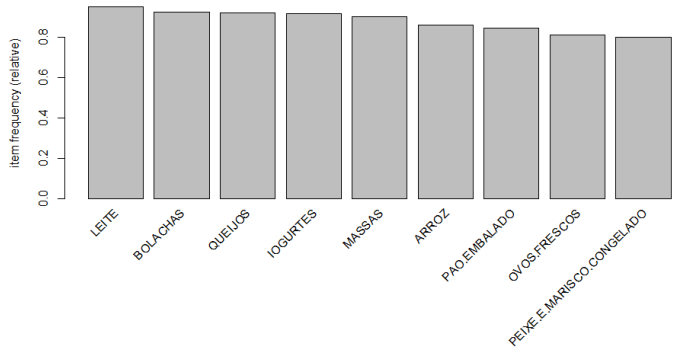

Figura 6.2 - Frequência de compra de artigos com *support* superior ou igual a 0,8

É possível observar que Leite é o produto mais frequente nos cabazes de compra dos portugueses, seguido da classe de produto Bolachas e Queijos. Considera-se que a maioria dos artigos apresentados nos cabazes dos consumidores se compreendem nos mercados dos bens essenciais, como Arroz e Massas, e dos produtos lácteos, como Leite, Queijos e Iogurtes.

Assim, e numa perspetiva de conseguir conclusões pertinentes, os artigos apresentados acima serão retirados, para a construção das regras de associação, pois, como a frequência é bastante elevada, as regras iriam incluir estes nove produtos alimentares como regras de associação fortes.

Numa primeira fase, através do algoritmo *Apriori*, pretende-se encontrar o grupo de artigos mais frequentes e derivar *itemsets* para cada nível da hierarquia do produto, consoante o valor pré-definido para o nível de *support*. Define-se o parâmetro *support*, retirando os produtos com frequência superior a 0,80, como mencionado. Não foi definido o valor mínimo, pois interessa verificar quais os *itemsets* mais frequentes, em cada nível da hierarquia. Após serem retirados os 9 artigos mais frequentes dos cabazes, o número de níveis hierárquicos diminuiu para três, contendo, agora, poucos *itemsets*.

Assim, e apesar dos resultados anteriormente apresentados serem favoráveis, quando são fixados os valores de *support* e *confidence* em 0,75, após retirados os 9 artigos mais frequentes, o número de níveis hierárquico diminui para apenas 1. Tornou-se, por isso, necessário utilizar outro método, com recurso às dez regras mais frequentes para cada nível hierárquico.

De seguida, na Tabela 6.3, é apresentado o conjunto de artigos com *support* inferior a 0,80, os segundos artigos mais frequentes. Neste conjunto de *itemsets*, apresentam-se dez para cada nível hierárquico existente, pois não se torna viável apresentar todas os *itemsets* existentes. É possível notar que os artigos apresentados pertencem, maioritariamente, aos mercados dos produtos enlatados, bebidas e bens essenciais.

Neste caso, nível hierárquico 1, o nível de *support* e *confidence*, entre os parâmetros, é igual, tendo sempre um *lift* de 1, em todos estes casos.

Tabela 6.3 - Os 10 *itemsets* mais frequentes com nível hierárquico 1

Itemsets	*Support*
{*Conservas de peixe*}	0,79
{*Fiambre*}	0,78
{*Vinhos*}	0,76
{*Água mineral*}	0,76
{*Margarina*}	0,75
{*Vegetais congelados*}	0,75
{*Açúcar*}	0,75
{*Batata frita*}	0,73
{*Vegetais em conserva*}	0,71
{*Produtos cárnicos fatiados*}	0,71

Através da Tabela 6.3, observam-se os 10 *itemsets* singulares mais frequentes, pertencentes ao nível hierárquico 1. É possível verificar que, dos 10 artigos mais frequentes, apenas dois pertencem ao mercado das bebidas: Água mineral e Vinhos. Por outro lado, os 8 artigos de comida apresentados não correspondem a produtos presentes no mercado dos bens essenciais, apresentando-se, em primeiro lugar, Conservas de peixe, seguidas de Fiambre, com 0,79 e 0,78, respetivamente.

Tabela 6.4 - Os 10 *itemsets* mais frequentes com nível hierárquico 2

Itemsets	*Support*
{*Fiambre, Conservas de peixe*}	0,64
{*Vinhos, Conservas de peixe*}	0,62
{*Margarina, Fiambre*}	0,62
{*Água mineral, Conservas de peixe*}	0,62
{*Produtos cárnicos fatiados, Fiambre*}	0,62
{*Vegetais em conserva, Conservas de peixe*}	0,62
{*Margarina, Conservas de peixe*}	0,61
{*Vinhos, Fiambre*}	0,61
{*Vegetais congelados, Fiambre*}	0,61
{*Vinhos, Água mineral*}	0,61

Na Tabela 6.4, observam-se os 10 *itemsets* mais frequentes, com nível hierárquico 2. Para uma correta aplicação de *itemsets*, que fossem oportunas de analisar, retirou-se as inversas às apresentadas, permanecendo as que tivessem maior valor de *confidence*.

Assim:

$Confidence(Fiambre, Conservas\ de\ Peixe) = \frac{0,64}{0,78} = 0,82$

enquanto que $Confidence(Conservas\ de\ Peixe, Fiambre) = \frac{0,64}{0,79} = 0,81$

O mesmo método foi utilizado para os restantes níveis hierárquicos apresentados abaixo.

Quando comparado o número de artigos presentes no nível hierárquico 1 com o 2, verifica-se que há 2 artigos que não estão presentes, Batatas fritas e Açúcar, sendo que os restantes 8 artigos estão presentes e formam os 10 *itemsets* mais frequentes. Deste modo, é possível averiguar que os artigos presentes se repetem na Tabela 6.4, fazendo com que os artigos Conservas de peixe e Fiambre façam parte de cinco *itemsets*, ou seja, metade das existentes. Não restam, assim, dúvidas de que o *itemset* $\{Fiambre, Conservas\ de\ peixe\}$ seja a que apresenta maior valor de *support*: 0,64.

Tabela 6.5 - Os 10 *itemsets* mais frequentes com nível hierárquico 3

Itemsets	*Support*
$\{Conservas\ de\ peixe, Margarina, Fiambre\}$	0,52
$\{Conservas\ de\ peixe, Produtos\ cárnicos\ fatiados, Fiambre\}$	0,51
$\{Vegetais\ em\ conserva, Fiambre, Conservas\ de\ peixe\}$	0,51
$\{Conservas\ de\ peixe, Vegetais\ congelados, Fiambre\}$	0,51
$\{Margarina, Produtos\ cárnicos\ fatiados, Fiambre\}$	0,50
$\{Conservas\ de\ peixe, Fiambre, Vinhos\}$	0,50
$\{Água\ mineral, Fiambre, Conservas\ de\ peixe\}$	0,50
$\{Água\ mineral, Vinhos, Conservas\ de\ peixe\}$	0,50
$\{Batata\ frita, Vegetais\ congelados, Fiambre\}$	0,50
$\{Batata\ frita, Fiambre, Conservas\ de\ peixe\}$	0,50

Com base nos resultados obtidos na Tabela 6.5, note-se que apenas são apresentadas *itemsets* em que a combinação de 3 artigos é única, ou seja, para a combinação de 3 artigos existem 3 *itemsets* disponíveis, considerando-se, por isso, a que tivesse o maior valor de *confidence*, de forma a que não existam *itemsets* repetidos em termos de compras simultâneas.

Analisando, em detalhe, o número de vezes que cada um dos produtos, Fiambre e Conservas de peixe, estão presentes nesta hierarquia, verifica-se que estes permanecem como artigos com número de presenças em *itemsets* existentes – nove e oito, respetivamente –, sendo que em sete das vezes estão presentes conjuntamente. Com isto, continuam a ser dos artigos de maior peso nos cabazes dos portugueses.

É, também, de notar, que, com o nível hierárquico 3, o artigo Batatas fritas aparece presente. Tal situação não aconteceu no nível anterior, fazendo parte das últimas duas do *itemset*. O *itemset* com maior *support* é $\{Conservas\ de\ peixe, Margarina, Fiambre\}$, com 0,52, considerando-se, desta forma, que quem compra Conservas de peixe, Margarina e Fiambre, estão presentes conjuntamente em 52% dos cabazes.

Tabela 6.6 - Os 10 *itemsets* mais frequentes com nível hierárquico 4

Itemsets	*Support*
{*Produtos de tomate, Vegetais em conserva, Fiambre, Conservas de peixe*}	0,43
{*Margarina, Fiambre, Produtos cárnicos fatiados , Conservas de peixe*}	0,43
{*Margarina, Vegetais em conserva, Fiambre, Conservas de peixe*}	0,42
{*Vegetais em conserva, Vegetais congelados, Fiambre, Conservas de peixe*}	0,42
{*Batata frita, Vegetais congelados, Fiambre, Conservas de peixe*}	0,42
{*Produtos de tomate, Margarina, Fiambre, Conservas de peixe*}	0,42
{*Produtos de Tomate, Fiambre, Produtos cárnicos fatiados, Conservas de peixe*}	0,41
{*Água mineral, Vinhos, Fiambre, Conservas de peixe*}	0,41
{*Produtos de tomate, Margarina, Produtos cárnicos fatiados, Fiambre*}	0,41
{*Vegetais em conserva, Fiambre, Produtos cárnicos fatiados, Conservas de peixe*}	0,41

Avaliando, de forma global, os resultados obtidos em todos os níveis hierárquicos, é possível verificar que o valor do *support* tem vindo a diminuir ao longo dos níveis, apesar de, em cada um deles, os valores se manterem constantes.

Através da Tabela 6.6, observa-se que o artigo Fiambre está presente em todos os *itemsets*, enquanto o artigo Conservas de peixe pertence a nove, das dez, *itemsets*. Sendo, esta, uma dupla de artigos que aparece em simultâneo desde o nível hierárquico 2, e ocupando o primeiro e segundo lugares, respetivamente, no nível hierárquico 1, era expectável que aparecessem, também, neste nível hierárquico, juntos. Por outro lado, os artigos, Água mineral e Vinhos, aparecem uma única vez juntos, assim como nos níveis anteriores.

Tendo em conta os resultados obtidos, nos diversos níveis hierárquicos acima apresentados, é essencial fixar um valor de *support*, que compreenda todos os *itemsets* com valor superior a 0,40, e o parâmetro *confidence*, em 0,65, de forma a obter-se uma boa variedade de regras de associação.

6.3. Regras de associação para cada nível hierárquico

Com a aplicação do algoritmo *Apriori* aos critérios definidos anteriormente, foram encontradas regras de associação para cada nível da hierarquia. Colocaram-se as regras de associação por ordem decrescente, no parâmetro *lift*, por forma a criar padrões de associação fortes. Note-se que a regra $\{A\} \rightarrow \{B\}$ tem o mesmo valor de *lift*, do que a regra $\{B\} \rightarrow \{A\}$, apesar de ambas terem valores de *confidence* diferentes, considerando-se, por isso, a regra que tivesse o valor mais elevado.

Foram analisados todos os níveis de hierarquia possíveis, 2, 3 e 4. Note-se que no nível hierárquico 1, o valor do parâmetro de *lift* é sempre 1, ou seja, irão obter-se exatamente os mesmos resultados anteriormente apresentados, não sendo, por isso, necessário expô-los novamente.

Tabela 6.7 - As 10 regras de associação com nível hierárquico 2

Regras	*Confidence*	*Lift*
$\{Produtos\ de\ tomate\} \rightarrow \{Vegetais\ em\ conserva\}$	0,80	1,13
$\{Produtos\ cárnicos\ fatiados\} \rightarrow \{Fiambre\}$	0,87	1,11
$\{Produto\ de\ tomate\} \rightarrow \{Produtos\ cárnicos\ fatiados\}$	0,78	1,11
$\{Produto\ de\ tomate\} \rightarrow \{Fiambre\}$	0,87	1,11
$\{Componentes\ de\ refeições\} \rightarrow \{Batata\ frita\}$	0,81	1,10
$\{Sal\} \rightarrow \{Especiarias\ e\ ervas\ aromáticas\}$	0,74	1,10
$\{Componentes\ de\ refeições\} \rightarrow \{Bebidas\ refrescantes\ s/\ gás\}$	0,77	1,10
$\{Produtos\ de\ tomate\} \rightarrow \{Margarina\}$	0,82	1,10
$\{Vegetais\ em\ conserva\} \rightarrow \{Conservas\ de\ peixe\}$	0,86	1,09
$\{Componentes\ de\ refeições\} \rightarrow \{Produtos\ de\ tomate\}$	0,73	1,09

A Tabela 6.7 indica, de uma forma geral, que o valor do parâmetro *lift* não apresenta grande variação, pertencendo ao intervalo $[1{,}09; 1{,}13]$. O parâmetro *confidence* apresenta-se, da mesma forma, com pouca variação de valores, para além dos seus baixos resultados.

Nas regras de associação para o nível hierárquico 2, a regra com a associação mais forte apresenta-se pela compra simultânea de Produtos de tomate e Vegetais em conserva. Porém, tal como é possível verificar, não se apresenta como a regra com maior nível de *confidence*, ou seja, em 80% dos cabazes de compras em que o consumidor comprar Produtos de tomate leva, também, Vegetais em conserva, sendo, estes, positivamente relacionados, devido ao valor de *lift* apresentado.

Tabela 6.8 - As 10 regras de associação com nível hierárquico 3

Regras	*Confidence*	*Lift*
$\{Vegetais\ em\ conserva, Produtos\ cárnicos\ fatiados\} \rightarrow \{Produtos\ de\ tomate\}$	0,81	1,21
$\{Vegetais\ em\ conserva, Fiambre\} \rightarrow \{Produtos\ de\ tomate\}$	0,81	1,21
$\{Conservas\ de\ peixe, Produtos\ de\ tomate\} \rightarrow \{Vegetais\ em\ conserva\}$	0,86	1,20
$\{Margarina, Produtos\ cárnicos\ fatiados\} \rightarrow \{Produtos\ de\ tomate\}$	0,80	1,19
$\{Margarina, Vegetais\ em\ conserva\} \rightarrow \{Produtos\ de\ tomate\}$	0,80	1,19
$\{Margarina, Fiambre\} \rightarrow \{Produtos\ de\ tomate\}$	0,79	1,18
$\{Batata\ frita, Produtos\ cárnicos\ fatiados\} \rightarrow \{Bebidas\ refrescantes\ s/\ gás\}$	0,82	1,18
$\{Bebidas\ refrescantes\ s/\ gás, Vegetais\ em\ conserva\} \rightarrow \{Produtos\ de\ tomate\}$	0,79	1,18
$\{Vinhos, Vegetais\ em\ conserva\} \rightarrow \{Produtos\ de\ tomate\}$	0,79	1,17
$\{Bebidas\ refrescantes\ s/\ gás, Produtos\ de\ tomate\} \rightarrow \{Produtos\ cárnicos\ fatiados\}$	0,83	1,17

Relativamente aos resultados obtidos na Tabela 6.8, sobre as regras de associação no nível hierárquico 3, apresentam-se valores baixos de *lift* e elevados de *confidence*. Os Produtos de tomate estão presentes em nove das dez regras existentes, seguindo-se os Vegetais em conserva, disponíveis em apenas seis regras.

É possível verificar que as duas primeiras regras apresentam valores iguais nos parâmetros existentes, sendo estas compostas igualmente por Vegetais em conserva e Produtos de tomate, variando apenas num produto – Produtos cárnicos fatiados ou Fiambre –, com um valor de *confidence* de 0,81 e de *lift*

1,21. Porém, esta simultaneidade mantém-se, uma vez que existem outras duas regras com os mesmos valores de parâmetros, diferenciando em apenas uma CP, que detêm, em paralelo, Margarina e Produtos de tomate, com o valores 0,80, de *confidence*, e 1,19, de *lift*.

Tabela 6.9 - As 10 regras de associação com nível hierárquico 4

Regras	*Confidence*	*Lift*
$\{Conservas\ de\ peixe, Vegetais\ em\ conserva, Fiambre\} \rightarrow \{Produtos\ de\ tomate\}$	0,84	1,25
$\{Margarina, Açúcar, Fiambre\} \rightarrow \{Produtos\ de\ tomate\}$	0,82	1,22
$\{Margarina, Fiambre, Produtos\ cárnicos\ fatiados\} \rightarrow \{Produtos\ de\ tomate\}$	0,82	1,22
$\{Bebidas\ refrescantes\ s/\ gás, Batata\ frita, Fiambre\} \rightarrow \{\ Produtos\ cárnicos\ fatiados\}$	0,85	1,21
$\{Conservas\ de\ peixe, Fiambre, Produtos\ cárnicos\ fatiados\} \rightarrow \{Produtos\ de\ tomate\}$	0,81	1,20
$\{Conservas\ de\ peixe, Margarina, Fiambre\} \rightarrow \{Produtos\ de\ tomate\}$	0,81	1,20
$\{Bebidas\ refrescantes\ s/gás, Conservas\ de\ peixe, Fiambre\} \rightarrow \{Produtos\ cárnicos\ fatiados\}$	0,83	1,17
$\{Conservas\ de\ peixe, Margarina, Fiambre\} \rightarrow \{Produtos\ cárnicos\ fatiados\}$	0,82	1,16
$\{Conservas\ de\ peixe, Vegetais\ congelados, Fiambre\} \rightarrow \{Vegetais\ em\ conserva\}$	0,83	1,16
$\{Conservas\ de\ peixe, Batata\ frita, Fiambre\} \rightarrow \{Bebidas\ refrescantes\ s/gás\}$	0,81	1,16

Por último, os resultados obtidos para o último nível hierárquico em análise, nível 4, são apresentados na Tabela 6.9. É possível observar que o maior valor de *lift*, o mais alto até agora em análise, mas mesmo assim bastante aquém dos valores usuais deste parâmetro, pertence à regra que indica que o consumidor no seu cabaz de compras, quando compra Conservas de peixe, Vegetais em Conserva e Fiambre, adquire, igualmente, em 80% das vezes, Produtos de tomate, apresentando uma positiva relação, entre os produtos mencionados, de 1,25.

Tal como na análise do nível hierárquico 3, também no nível 4 existem dois conjuntos de regras que apresentam os mesmos valores, nos parâmetros em análise. O primeiro conjunto pertence às regras que se encontram nas posições 2 e 3, que englobam, em ambas, as CPs de Margarina e Produtos de tomate, com um total de 0,82, de *confidence*, e 1,22, de *lift*. O segundo conjunto diz respeito às regras 5 e 6, que incluem, simultaneamente, os produtos Conservas de peixe e Produtos de tomate e apresentam um valor de 0,81, de *confidence*, e 1,20, de *lift*.

Em relação à análise desenvolvida sobre todos os níveis hierárquicos existentes, observa-se que à medida que o nível hierárquico aumenta, maior é também o nível de *lift* da regra de associação, ou seja, maiores as fortes relações de dependência entre as CPs apresentadas. Relativamente ao parâmetro *confidence*, nada se pode concluir, pois são apresentados valores idênticos, em todos os níveis da hierarquia.

Algumas das regras anteriormente apresentadas, nos vários níveis hierárquicos, podem ter sido potencializadas pela forte atividade promocional, existente em Portugal. Porém, e de forma a retirar conclusões mais concretas sobre esta temática, seria necessária uma análise complementar relativamente à sazonalidade das associações presentes, bem como verificar se as regras entre os artigos mantêm os mesmos valores, nos parâmetros de *confidence* e de *lift*, ao longo do ano.

6.4. Regras de associação para *lift* inferior ou igual a 1

Até agora, foram apresentadas as regras mais fortes, em todos os níveis de hierarquia existentes. Nesta secção, irá analisar-se as regras que tenham o parâmetro *lift* inferior ou igual a 1, por representarem categorias que estão negativamente relacionadas ou com fraca relação, isto é, independentes. Estes dois cenários serão analisados de uma forma geral, sem distinguir as regras de associação por nível hierárquico.

Tabela 6.10 - As 10 regras de associação com valor de *lift* igual a 1

Regras	*Confidence*
$\{Açúcar, Especiarias\ e\ ervas\ aromáticas\} \rightarrow \{Água\ mineral\}$	0,76
$\{Bacalhau\} \rightarrow \{Vegetais\ congelados\}$	0,75
$\{Conservas\ de\ peixe, Sal\} \rightarrow \{Vegetais\ congelados\}$	0,75
$\{Manteiga, Fiambre\} \rightarrow \{Margarina\}$	0,75
$\{Conservas\ de\ peixe, Componentes\ de\ refeições\} \rightarrow \{Açúcar\}$	0,75
$\{Água\ mineral, Batata\ frita\} \rightarrow \{Açúcar\}$	0,75
$\{Conservas\ de\ peixe, Açúcar\} \rightarrow \{Vegetais\ congelados\}$	0,75
$\{Azeite\} \rightarrow \{Produtos\ cárnicos\ fatiados\}$	0,71
$\{Sal\} \rightarrow \{Bebidas\ refrescantes\ s/\ gás\}$	0,70
$\{Bolos\ embalados\} \rightarrow \{Especiarias\ e\ ervas\ aromáticas\}$	0,68

Na Tabela 6.10, apresentam-se todas as regras de associação pertencentes aos níveis hierárquicos 2 e 3. Não foram considerados os resultados do nível hierárquico 1, pois as regras apenas contêm um único artigo, tendo sempre o parâmetro *lift* igual a 1. Relativamente ao nível hierárquico 4, não foram encontrados resultados em que o parâmetro fosse igual ao valor pretendido.

É possível observar que as 10 regras, com o parâmetro fixo em 1, apresentam um nível de *confidence* bastante aceitável, apontando o conjunto de regras que representam produtos independentes. Estas regras não têm interesse, porque a sua associação é aleatória e, por isso, nada se pode concluir.

Tabela 6.11 - As 10 regras de associação com menor *lift*

Regras	*Confidence*	*Lift*
$\{Manteiga\} \rightarrow \{Margarina\}$	0,71	0,94
$\{Iogurtes, Manteiga\} \rightarrow \{Margarina\}$	0,72	0,95
$\{Conservas\ de\ peixe, Manteiga\} \rightarrow \{Margarina\}$	0,72	0,96
$\{Componentes\ de\ refeições\} \rightarrow \{Cafés\ torrados\}$	0,65	0,98
$\{Sal\} \rightarrow \{Componentes\ de\ refeições\}$	0,66	0,98
$\{Cafés\ torrados\} \rightarrow \{Componentes\ de\ refeições\}$	0,66	0,98
$\{Cereais\ p/\ pequeno\ almoço\} \rightarrow \{Especiarias\ e\ ervas\ aromáticas\}$	0,67	0,98
$\{Componentes\ de\ refeições\} \rightarrow \{Especiarias\ e\ ervas\ aromáticas\}$	0,67	0,99
$\{Frutos\ secos\ embalados\} \rightarrow \{Bebidas\ refrescantes\ s/\ gás\}$	0,69	0,99

Depois de apresentadas as regras de associação de artigos independentes, analisam-se as regras que detêm um parâmetro *lift* inferior a 1, ou seja, classes de produto que estão negativamente relacionadas. Através da Tabela 6.11, observam-se as 10 regras que apresentam os menores valores de parâmetro.

Neste tipo de casos, é recorrente a existência de *canibalismo* entre artigos, um conhecido problema de Marketing, que ocorre quando a venda de um dos produtos de determinada empresa reduz as vendas de outros. Este tipo de situações acontece aquando da entrada de um novo produto, da mesma família, no mercado (Kotler, 2000)[20]. Neste caso, em específico, não é possível retirar este tipo de conclusões, uma vez que se está perante uma análise de CPs e não de artigos.

A regra que apresenta a relação mais negativa entre duas CPs, Manteiga e Margarina, tem um valor de parâmetro *lift* de 0,94, seguida de Iogurtes, Manteiga e Margarina, com um total de 0,95 e, em terceiro lugar, Conservas de peixe, Manteiga e Margarina, com um resultado de 0,96. Conclui-se, então, que não existem muitas regras de associação com relação negativa, visto que, nas regras 9 e 10, os resultados estão muito próximos de 1, com um valor de 0,99.

[20] Hoje em dia, a área de produtos eletrónicos é a mais permeável a este tipo de situações. Um bom exemplo disso, pode ser encontrado em https://news4c.com/ipad-mini-series-discontinued-the-curious-case-of-ipad-mini-5/ (consultado em 2 de janeiro de 2017).

Conclusões

Neste trabalho, analisou-se o consumo de bens alimentares, em famílias portuguesas. Para isso, foram utilizados dados reais, fornecidos pela empresa líder mundial de estudos de mercado, *The Nielsen Company*. Desta forma, o tema de investigação, *Data Mining* e Identificação de Padrões, foi analisado em duas vertentes. A primeira, através de metodologias de *clustering*, segmentando-se os lares por presença de produto no cabaz de compras e quantidade comprada. A segunda, através de *market basket analysis*, com especial foco nas regras de associação.

A realização da análise de segmentação de perfis de consumo, a partir da formação de *clusters* naturais, teve como objetivo a criação de grupos de famílias que compram o mesmo tipo de produtos. Os dados disponíveis, por presença por tipo de produto, foram segmentados utilizando coeficientes, modelos e os métodos de corte existentes. Porém, os *outputs* recebidos não foram os esperados. Desse modo, não foi possivel averiguar qual o critério que melhor resultou na segmentação dos lares. Os *clusters* tinham maioritariamente o mesmo comportamento, ou seja, compravam mais ou menos os mesmos produtos, não existindo, dessa forma, grandes dissemelhanças. Em termos de características sociodemográficas dos grupos segmentados, obteve-se exatamente o mesmo cenário – valores idênticos e sem grandes oscilações entre si.

Assim, tornou-se necessária a realização de uma outra análise, tendo em conta a variedade de produtos comprados. Verificou-se, aí, que a formação de *clusters*, a partir da segmentação por quantidade comprada, no cabaz de compras, continua a não apresentar bons resultados, quando analisadas as características sócio demográficas das 447 famílias em estudo. Assim, não foi possível compreender o critério utilizado para a segmentação dos lares a partir das duas distâncias construídas, *my.dist* e *my.dist2*, e dos dois índices utilizados, *NbClust* e *PAM*, pelo facto de não se identificarem *clusters* homogéneos, em praticamente toda a análise.

Destaca-se, contudo, a divisão dos dados em 3 *clusters*, utilizando o *package NbClust.* Com isso, foram obtidos *clusters* homogéneos e as diferenças, em relação à frequência de compra e variedade de CPs no

cabaz de compras, são significativas. Com estes dois indicadores, foi possível identificar algumas dissemelhanças nos *clusters*, não se conseguido, porém, e como na análise anterior, definir o melhor cenário a aplicar aos dados.

Para a concretização de estratégias de marketing, é necessária a tomada de decisões. Desse modo, as características analisadas não são suficientes para daí se tirarem conclusões robustas, com o objetivo de as conseguir aplicar em problemas atuais, existentes no mercado de bens de grande consumo. Não era, contudo, esperado que dados de 3,5 meses fossem suficientes para analisar tendências e perfis de consumo, do mercado português. Esta seria, sim, uma forma de as empresas do setor conseguirem obter alguma vantagem em relação à concorrência, uma vez que teriam à sua disposição uma análise detalhada de perfis de consumo, que embora não sendo representativa da população portuguesa, daria, certamente, indicadores que pudessem melhorar o mercado.

Relativamente à última análise, realizada neste trabalho, *market basket analysis*, esta revela ser uma área de estudo de grande importância e finalidade prática para as empresas do setor, por permitir encontrar relações de dependência entre artigos. Desta forma, ajuda igualmente a descrever o comportamento dos consumidores portugueses, retirando partido dos padrões de compra de bens de grande consumo. Isto permite às empresas a obtenção de vantagem competitiva, face à concorrência.

A dimensão da base de dados, bem como a classificação dos artigos em classes de produto gerais, impediu o alcance de conclusões mais profundas acerca das relações entre os artigos dentro da mesma família. Para além disso, a análise das regras de associação tornou-se mais difícil, uma vez que o software utilizado, por si só, não elimina regras redundantes, tendo sido estas desconsideradas, manualmente.

Em Portugal, no período em análise, os artigos mais frequentes, apresentados nos cabazes dos consumidores, compreendem-se nos mercados dos bens essenciais, como Arroz e Massas, e dos produtos lácteos, como Leite, Queijos e Iogurtes, que estão presentes em mais de 80% dos cabazes de compras.

Os artigos mais frequentes formaram regras de associação mais frequentes. Desse modo, e após retirados os produtos que permaneciam em mais de 80% dos cestos de compras dos portugueses e eliminadas as regras redundantes, verificou-se que, depois disso, Fiambre e Conservas de peixe são os produtos mais comercializados, em todos os níveis hierárquicos. Para além disso, verificou-se, ainda, que nenhum dos artigos apresentados nos quatro níveis hierárquicos é considerado saudável, nem corresponde a qualquer artigo presente no mercado dos bens essenciais.

Foram analisados, também, todos os níveis de hierarquia possíveis – 2, 3 e 4 –, tendo-se averiguado que a regra com maior valor de parâmetro de *lift* pertence à regra que indica que o consumidor que se desloca ao supermercado para comprar Conservas de peixe, Vegetais em Conserva e Fiambre compra igualmente, em 80% das vezes, Produtos de tomate, apresentando uma positiva relação, entre os produtos mencionados, de 1,25.

Através da análise das regras que detêm um parâmetro *lift* inferior a 1, ou seja, classes de produto que estão negativamente relacionadas, é possível concluir que a relação mais negativa, entre dois artigos, pertence às CPs Manteiga e Margarina, com um valor de parâmetro *lift* de 0.94, seguida de Iogurtes, Manteiga e Margarina, com um total de 0.95. Este é um cenário que está intimamente relacionado com

a existência de canibalismo entre artigos, um conhecido problema de Marketing, que ocorre quando a venda de um dos produtos de determinada empresa reduz as vendas de outros.

As conclusões retiradas desta análise podem ser utilizadas, por empresas de Marketing e logística, de forma a influenciar a visibilidade dos artigos e classes de produtos mais vendidos, no supermercado. A principal estratégia passaria pela colocação de artigos com relação positiva no mesmo linear. Por outro lado, afastar conjuntos de artigos que pertençam a regras de elevado valor, no parâmetro *confidence*, seria outra das opções, implicando que o consumidor percorra mais corredores e seja obrigado a observar mais produtos. Para além disso, no caso de promoção de artigos em ações de cross-selling e promoções, em situações em que o nível de *confidence* é bastante elevado – quando o antecedente da regra atinge as vendas esperadas e o consequente não –, poderão ser propostos alguns tipos de estratégia em que, na compra do antecedente, se oferece uma promoção no consequente da regra, potenciando, assim, a venda deste último.

Atualmente, os métodos e estratégias de Marketing encontram-se bastante desenvolvidos, em Portugal. São considerados meios rápidos e baratos e permitem, a qualquer empresa de média ou grande dimensão, conseguir obter todo o tipo de informação que precisa, de forma a aumentar a vantagem competitiva em relação à concorrência. Desse modo, o principal, e comum, objetivo é antecipar as preferências dos consumidores, com base nas preferências ou histórico de compras, como forma de ir ao encontro dos seus comportamentos. Só assim, as empresas conseguirão obter vantagem face às restantes congéneres do setor em que operam e, consequentemente, aumentar o volume de vendas e o respetivo lucro.

Referências Bibliográficas

Armstrong, G., & Kotler, P. (2011). *Principles of Marketing*. New Jersey: Pearson Prentice Hall.

Barbosa, A. P. (2017, janeiro-junho). Que tendências irão marcar o ano de 2017?. *Código 560*, (28), 36-37.

Branco, J. A. (2004). *Uma Introdução à Análise de Clusters*. Évora: Sociedade Portuguesa de Estatística.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*(6), 1-36. Consultado em 29 de junho de 2017
em http://www.jstatsoft.org/v61/i06/

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Chichester: John Wiley & Sons, Ltd.

Gama, J., Carvalho, A. P. L., Faceli, K., Lorena, A. C., & Oliveira, M. (2012). *Extração do Conhecimento de Dados - Data Mining*. Lisboa: Edições Sílabo.

Hahsler, M., & Hornik, K. (2007). New probabilistic interest measures for association rules. *Intelligent Data Analysis*, *11*(5), 437-455.

Hahsler, M., Grun, B., & Hornik, K. (2005). A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, *14*(15), 1-25.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann Publishers.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Cambridge: A Bradford Book.

Instituto Nacional de Estatística, & Pordata. (2015). *Dimensão média das famílias segundo os Censos*. Consultado em 17 de maio de 2017 em http://www.pordata.pt/Portugal/Dimens%C3%A3o+m%C3%A9dia+das+fam%C3%ADlias+segundo+os+Censos++-908

Instituto Nacional de Estatística. (2011). *Censos 2011 - Resultados Provisórios*. Consultado em 19 de maio de 2017 em https://pt.scribd.com/document/212858142/Censos2011-ResultadosProvisorios

Instituto Nacional de Estatística. (2017). *Famílias clássicas na população residente por Tipo de família clássica. Inquérito ao emprego*. Consultado em 15 de maio de 2017 em https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0007861&contexto=bd&selTab=tab2

Johnson, R. A., & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. New Jersey: John Wiley & Sons.

Kotler, P. (2000). *Marketing Management, Millenium Edition*. New Jersey: Pearson Prentice Hall.

Lavrenko, V. [Victor Lavrenko]. (2015, setembro 14). *IAML19.5 Single-link, complete-link, Ward's method* [Ficheiro de vídeo]. Consultado em 12 de março de 2017 em https://youtu.be/vg1w5ZUF5lA?t=491

Marktest. (2010). *Atlas Social de Portugal 2010*. Consultado em 20 de março de 2017 em http://www.marktest.com/wap/private/images/logos/Folheto%20AtlasSocial.pdf

Microstrategy. (2003). *Business Intelligence in the Retail Industry*.

Milligan, G. W., & Cooper, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika, 50*(2), 159-179.

Nielsen Portugal. (2016a). *Consumidores portugueses valorizam produtos Premium*. Consultado em *14 de maio de 2016* em http://www.nielsen.com/pt/pt/press-room/2016/consumidores-portugueses-valorizam-produtos-premium.html

Nielsen Portugal. (2016b). *O Consumidor está a mudar? #Nielsen 360º apresenta 2016 com boas perspectivas*. Consultado em 15 de maio de 2017 em http://www.nielsen.com/pt/pt/press-room/2016/the-consumer-is-changing-nielsen-360-shows-in-2016-with-good-prospects.html

Nielsen Portugal. (2017a). *Confiança dos Portugueses atinge o valor mais alto de sempre*. Consultado em 17 de maio de 2017 em http://www.nielsen.com/pt/pt/press-room/2017/confidence-of-the-portuguese-reaches-the-highest-value-ever.html

Nielsen Portugal. (2017b). *Consumidores portugueses aderem às compras online*. Consultado em 12 de maio de 2017 em http://www.nielsen.com/pt/pt/press-room/2017/portuguese-consumers-join-online-shopping.html

O Gang dos Frescos está de volta! [Versão online], *Activa*. (2015). Consultado em 1 de maio de 2017 em http://activa.sapo.pt/saude-e-beleza/2015-10-06-O-Gang-dos-Frescos-esta-de-volta-1

Raeder, T., & Chawla, N. V. (2011). Market Basket Analysis with Networks. *Analysis and Mining*, *1*(2), 97-113.

Reis, E. (2001). *Estatística Multivariada Aplicada*. Lisboa: Edições Sílabo.

Rodrigues, M., Gama, J., & Ferreira, C. A. (2012). Identifying Relationships in Transactional Data. In J. Pavón, N. D. Duque-Méndez & R. Fuentes-Fernández (Eds.). *Advances in Artificial Intelligence - IBERAMIA 2012* (Vol. 14, Chap. 15, pp. 81-90). Berlin: Springer Heidelberg.

Salvador, A. B., & Campomar, M. C. (2014). Segmentação e posicionamento: o coração do plano de marketing. *Inovcom: Revista Brasileira de Iniciação Científica em Comunicação*, *6*(1), 41-50. Consultado em 29 de agosto de 2017 em http://www.portcom.intercom.org.br/revistas/index.php/inovcom/article/viewFile/1852/1674

Silva, A. R. (2016). *Grandes superfícies multadas em 920 mil euros por venderem com prejuízo*. Consultado em 29 de maio de 2017 em https://www.publico.pt/2016/07/14/economia/noticia/supermercados-multados-em-920-mil-euros-por-promocoes-agressivas-1738153

Timm, N. H. (2002). *Applied Multivariate Analysis*. New York: Springer-Verlag.

Ulas, M. (1999). *Market Basket Analysis for Data Mining*. Master Dissertation, Bo˘gaziçi University, Boğaziçi University, Istanbul, Turkey. Consultado em 29 de maio de 2017 em https://www.cmpe.boun.edu.tr/~ulas/msthesis.pdf

Wei, C. P., Lee, Y. H., & Hsu, C. M. (2003). Empirical comparison of fast partitioning-based clustering algorithms for large data sets. *Expert Systems with applications*, *24*(4), 351-363.

Zaki, M. (2000). *Generating Non-Redundant Association Rules*. Consultado em 29 de maio de 2017 em http://www.cs.tau.ac.il/~fiat/dmsem03/Generating%20Non-Redundant%20Association%20Rules%20-%202000.pdf

Anexos

Anexo A - Áreas *Nielsen*

Área I - Grande Lisboa: Concelhos de Lisboa, Odivelas, Loures, Amadora, Oeiras, Cascais, Sintra, Almada, Barreiro e Seixal.
Área II - Grande Porto: Concelhos do Porto, Matosinhos, Gaia, Maia, Gondomar e Valongo.
Área III Norte - Litoral Norte: Distritos: Viana Castelo, Braga, Aveiro e Resto do Distrito do Porto. Concelhos de Coimbra: Coimbra, Cantanhede, Condeixa, Figueira da Foz, Mira, Soure e Montemor-o-Velho.
Área III Sul - Litoral Sul: Distrito Leiria: Concelhos de Leiria, Alcobaça, Batalha, Bombarral, Caldas Rainha, Marinha Grande, Nazaré, Óbidos, Peniche, Pombal e Porto de Mós. Distrito Santarém: Concelhos de Santarém, Cartaxo e Rio Maior. Distrito Lisboa: Resto do Distrito de Lisboa. Distrito Setúbal: Concelhos de Setúbal, Montijo, Alcochete, Moita, Palmela e Sesimbra.
Área IV - Interior Norte: Distritos de Bragança, Vila Real, Guarda, Viseu, Castelo Branco. Resto dos Distritos de Coimbra e Leiria.
Área V - Interior Sul: Distritos de Portalegre, Évora, Beja e Faro. Resto dos Distritos de Santarém e Setúbal.

Anexo B - Tipos de família *Nielsen*

Pré-Famílias: Agregados com apenas um membro com idade inferior a 35 anos. Agregados de 2 ou mais membros, em que a dona de casa tem menos de 35 anos, e sem elementos abaixo dos 18 anos.
Famílias Novas: Agregados com crianças em que todas têm menos de 6 anos.
Famílias em Desenvolvimento: Agregados com crianças ou jovens (0-17 anos), mas nem todas as crianças têm menos de 6 anos ou mais de 10 anos (ou seja, não incluídas nas Famílias Novas ou Famílias Estáveis)

Famílias Estáveis: Agregados com todas as crianças/jovens acima dos 10 anos.
Famílias Maduras: Agregados com apenas um membro com idade entre os 35 e os 54 anos inclusive. Agregados de 2 ou mais membros em que a dona de casa tem idade entre os 35 e os 54 anos, sem crianças ou jovens com idade inferior a 18 anos.
Seniores: Agregados com mais de um membro, com a dona de casa com idade superior ou igual a 55 anos, e sem crianças ou jovens de idade inferior a 18 anos. Agregados com apenas um membro com idade superior ou igual a 55 anos.